Importance de la temporalité dans les phénomènes de propagation. Une illustration sur des échanges d'animaux d'élevage

Aurore Payen

THÈSE DE DOCTORAT DE SORBONNE UNIVERSITÉ

Spécialité **Informatique**
École doctorale Informatique, Télécommunications et Électronique (Paris)

Présentée par
Aurore Payen

Pour obtenir le grade de

DOCTEUR de SORBONNE UNIVERSITÉ

IMPORTANCE DE LA TEMPORALITÉ DANS LES PHÉNOMÈNES DE PROPAGATION.
Une illustration sur des échanges d'animaux d'élevage

Soutenue le 8 octobre 2018 devant le jury composé de :

Rapporteurs :	
M. Renaud Lambiotte	Professeur, Oxford University
M. Camille Roth	Professeur, SciencesPo Paris
Examinateurs	
M. Christophe Cambier	Maître de conférences, Sorbonne Université/IRD
M. Pascal Crépey	Professeur, EHESP
M. Elisabeta Vergu	Directeur de recherches, Inra
Directeurs	
M. Matthieu Latapy	Directeur de recherches, CNRS
M. Lionel Tabourier	Maître de conférences, Sorbonne Université

Remerciements

Je remercie tout d'abord le ministère de l'Agriculture et de l'Alimentation pour l'opportunité offerte de développer mes compétences dans le cadre de ce projet de thèse.

Je suis particulièrement reconnaissante envers mes rapporteurs Renaud Lambiotte et Camille Roth, ainsi que Christophe Cambier, Pascal Crépey, et Elisabeta Vergu, qui ont accepté d'être membres du jury.

Je tiens à remercier Matthieu et Lionel pour leur suivi tout au long de cette thèse : pour leur disponibilité et les points très réguliers que nous avons pu faire, pour leur accompagnement dans l'acquisition des compétences nécessaires à l'obtention de ce doctorat, et pour leur bienveillance.

Mes remerciements également à tout le reste de l'équipe, présente et passée, permanents, post-doc, ingé et thésards, pour la bonne ambiance qui a toujours régnée, pour les bons souvenirs qui ont rempli ces 3 années et contribué à rendre cette expérience inoubliable : Clémence, Fabien, Max, Robin, Keun-woo, Élie, Rémy C., Pédro, Louisa, Noé, Thibaud, Tiphaine, Raphaël, Fréd, Hong-lan, Léonard, Léo, Marwan, Rémy P., Audrey.

J'ai enfin une pensée emplie d'affection pour ma famille et mes amis, pour tout ce qu'ils m'ont apporté tout au long de cette thèse.

Table des matières

Introduction

Dans notre monde globalisé, les échanges commerciaux entre exploitations agricoles favorisent la diffusion à grande échelle des maladies. L'enjeu est non seulement économique, de par les répercussions fortes sur les marchés de produits d'origine animale en cas de crise, mais également de santé publique, de nombreuses maladies animales étant transmissibles à l'homme (comme la tuberculose bovine). La traçabilité des animaux devient alors une question de plus en plus importante pour retrouver les foyers infectieux, et lutter contre la propagation des maladies. On observe donc à l'échelle mondiale la mise en place de systèmes d'enregistrement des échanges d'animaux d'élevage, et notamment des bovins : selon les pays, cet enregistrement peut être soumis au volontariat des exploitants, comme au Canada, ou faire l'objet d'une réglementation nationale, comme en Uruguay, voire internationale, comme en Union européenne.

Depuis les années 2000, les échanges d'animaux doivent obligatoirement être tracés dans chaque pays de l'union européenne. En France, la base de données nationale d'identification (BDNI) correspond à la mise en place de ce dispositif. Il existe de plus une coopération entre pays pour permettre le suivi entre les états membres, en transmettant les données des animaux importés et exportés. Le développement de ce type de données et leur accessibilité a permis l'émergence d'études sur leur organisation et leur dynamique. Les outils et modèles développés pour l'étude des réseaux sociaux ont été adaptés à l'étude de ces données.

Cette thèse s'inscrit dans ce cadre : son apport réside dans l'utilisation de mesures et de modèles intégrant l'information temporelle sur les échanges d'animaux. En effet, le développement des réseaux temporels est relativement récent, et peu d'études les ont à ce jour appliqué aux réseaux d'échanges d'animaux. L'objectif est donc double, participer au développement des outils d'analyse des réseaux temporels, et en déduire des pistes de développement de mesures de surveillance et de lutte contre la propagation des maladies entre exploitations.

Dans le premier chapitre, nous reviendrons sur les mesures et modèles définis dans la littérature pour étudier les réseaux d'élevage, et modéliser la propagation de maladies. Ensuite, nous décrirons dans le deuxième chapitre les données d'enregistrement des animaux et de leurs mouvements auxquelles nous avons eu accès. Dans le troisième chapitre, nous évaluerons l'impact de modèles simples de propagation de maladies sur ces données, en prenant en compte l'information temporelle sur les échanges d'animaux. Puis, nous simulerons dans le quatrième chapitre des stratégies de lutte contre la propagation, et nous montrerons comment intégrer l'information temporelle sur les échanges permet d'améliorer significativement leur efficacité. Enfin, le cinquième chapitre présente l'étude de l'accessibilité au cours du temps des exploitations par les propagations. Ceci nous est rendu possible par

l'utilisation de l'information temporelle des données. Nous décrirons comment modéliser cette dynamique. Nous conclurons ce travail de thèse par différentes perspectives : notamment, nous proposerons des pistes d'amélioration de la modélisation de la dynamique de propagation, et d'étude du lien entre les caractéristiques du réseau et la dynamique de diffusion.

Des résumés de sections vous seront proposés, afin de permettre une lecture rapide des parties de résultats.

Chapitre 1

État de l'art

Sommaire

Dans ce chapitre, nous présentons les motivations aboutissant à l'utilisation de réseaux temporels pour l'étude des échanges d'animaux d'élevages. Nous présentons leurs utilisations dans le cadre de la modélisation de diffusion de maladies, et les notions essentielles à leur étude.

1.1 Éléments contextuels

1.1.1 La santé des bovins, enjeu majeur pour la production alimentaire

La santé humaine est étroitement liée à la santé animale et à l'environnement. Par exemple, les animaux sont à l'origine de 60% des maladies humaines infectieuses. L'aug-

mentation des flux à l'échelle mondiale favorise de plus leur propagation. Une approche intégrée de la santé a donc été développée, afin de renforcer la coordination entre les questions de santé humaine, animale, et l'environnement. Cette approche vise notamment à développer et améliorer la prévention, la surveillance et l'efficacité de la lutte contre les maladies animales [Direction générale de la mondialisation, 2011]. Les zoonoses présentées comme majeures par l'organisation mondiale de la santé sont pour la plupart portées par les animaux d'élevage (par exemple la brucellose, la maladie du charbon, ou les fièvres hémorragiques). Un nombre important d'entre elles touche plus particulièrement les bovins, comme c'est le cas de la tuberculose bovine ou de l'encéphalopathie spongiforme bovine (ESB).

La viande bovine est la deuxième viande la plus exportée dans le monde en termes de quantité avec 7,8 millions de tonnes équivalent carcasse en 2011 [FranceAgriMer, 2012]. Les crises sanitaires sont donc un enjeu majeur pour les pays producteurs comme la France, avec de fortes répercussions économiques sur les filières bovines en cas d'embargo. La détection précoce et la gestion rapide des crises sanitaires sont donc essentielles pour en limiter les conséquences économiques en récupérant rapidement le statut indemne.

De nombreuses voies de propagation sont susceptibles de contaminer un élevage : les contacts entre troupeaux au pâturage, avec la faune sauvage, la propagation aérienne des germes, par introduction d'un animal contaminé dans un cheptel, etc. Cependant, le rôle de ces voies de propagation varie selon l'échelle de l'étude, et la maladie considérée. Par exemple, [Green et al., 2006] montrent sur le réseau d'élevage de Grande-Bretagne, que les mouvements des animaux par introduction de nouveaux individus dans un élevage, notamment par achat, expliquent la propagation de la fièvre aphteuse à l'échelle nationale, alors que la transmission locale de la maladie, par exemple par contact au pâturage, a surtout un impact sur l'ampleur des crises. Les échanges d'animaux entre exploitations sont souvent retenus comme voie prépondérante dans la diffusion de certaines maladies, comme par exemple dans le cas de la fièvre aphteuse [Dubé et al., 2009], la tuberculose bovine et la fièvre catarrhale [Ensoy et al., 2014]. Si ces échanges ne constituent pas la seule voie de contagion pour les exploitations, ils jouent tout de même un rôle central dans la diffusion de maladies. Comprendre leur structure et leur temporalité est donc essentiel.

En cas de crise sanitaire, l'enjeu pour les autorités sanitaires est de réussir à contenir et éradiquer la maladie le plus rapidement possible. Il en découle par exemple la surveillance ou l'abattage préventif des troupeaux dans lesquels un animal originaire d'un élevage infecté a été introduit. Modéliser les échanges d'animaux entre élevages permettrait, d'une part, une meilleure gestion en amont, en ciblant plus efficacement les troupeaux à surveiller et en testant différentes mesures préventives au moyen de simulations, et d'autre part une meilleure gestion de crise, en testant l'efficacité des différents moyens de lutte, toujours au moyen de simulations.

1.1.2 Apports des approches réseaux

Depuis la crise d'encéphalopathie spongiforme bovine (ESB) de 1996, où l'importation de bovins britanniques avait entraîné la propagation de la maladie en France et en Europe, l'Union européenne a rendu obligatoire la création de bases de données, enregistrant chaque bovin et ses changements successifs d'exploitation [EU, 2000]. Ces bases permettent de retracer les échanges d'animaux entre exploitations, et donc la voie de propagation par introduction d'animaux malades. Autrement dit, elles fournissent la liste des échanges effectués entre toutes les paires d'exploitations. L'utilisation des approches réseaux pour modéliser les échanges d'animaux d'élevage entre exploitations est une solution de plus en plus utilisée dans le domaine de l'épidémiologie. Par approches réseaux, nous désignons l'ensemble des approches permettant l'étude de l'existence de contacts entre acteurs, étant ainsi centrées sur la structure des interactions plus que sur la nature des agents. Selon cette définition, elles englobent à la fois les réseaux temporels et les graphes (voir partie suivante).

• **Apports pour la description**

Étudier la liste des échanges par ces approches permet de décrire leur structure. Par exemple, [Robinson et al., 2007] étudie l'impact sur le réseau d'échanges de la mise en place en Grande-Bretagne d'une période d'immobilisation de l'animal acheté sur l'exploitation de l'acquéreur. Les auteurs montrent qu'au lieu de diminuer les risques de contamination en diminuant le nombre de changements d'exploitations que peut effectuer un animal, le réseau s'est réorganisé et offre maintenant un nombre plus grand de voies de propagation possibles pour la diffusion d'une maladie. Les exploitations seraient donc exposées à un risque accru de contamination, en cas de crise. Cette étude insiste donc sur le fait que toute mise en œuvre d'une mesure peut avoir des répercussions non prévues sur la structure du réseau et sur sa vulnérabilité à la propagation de pathogènes.

En outre, utiliser une représentation réseau des échanges permet de caractériser le comportement des exploitations qui le composent, comme nous le verrons plus en détails dans la partie 1.2.3 et au chapitre 2. Cette connaissance des acteurs permet d'évaluer leur risque d'exposition à la diffusion de maladies, et d'adapter en conséquence les méthodes de surveillance et de lutte. Par exemple, une exploitation échangeant un nombre très important d'animaux entraîne potentiellement un risque plus élevé d'infection que pour les autres exploitations. En s'inspirant de travaux sur la surveillance basée sur le risque, comme dans [Cameron, 2012], l'identification grâce aux approches réseaux des exploitations les plus exposées pourrait permettre d'améliorer les méthodes de surveillance. En effet, cela permettrait de diminuer la fréquence des tests sur les exploitations exposées à un risque faible d'infection, et de renforcer le contrôle des exploitations à risque. De plus, en cas de crise sanitaire, les approches réseaux pourraient permettre à terme de détecter quelles exploitations seraient à cibler en priorité par les mesures de lutte, afin de limiter les impacts

économiques de l'épizootie (perte du statut indemne, perte de production...).

• **Apports pour la simulation**
Un réseau, avec ou sans information temporelle, peut être utilisé dans la simulation de diffusions. En effet, il est le support de la dynamique des interaction et permet de décrire leur structure, informations essentielles pour modéliser la diffusion de maladies : les transferts d'animaux qui ont réellement eu lieu pourront être utilisés pour simuler la propagation. Si de nombreuses études modélisent le cas spécifique des maladies animales, comme nous le verrons en partie 1.3, ce pan de la littérature bénéficie également des outils méthodologiques développés pour l'étude des maladies humaines (voir par exemple [Morris and Kretzschmar, 1997] pour l'étude de la propagation du VIH), ou pour l'étude des phénomènes diffusifs en général, comme nous y reviendrons dans la partie suivante.
De plus, [Keeling and Eames, 2005] souligne l'importance d'utiliser une approche réseau pour simuler précisément des mesures de lutte. En effet, la simulation d'épisodes infectieux rend possible l'évaluation des effets des méthodes de lutte (abattage préventif, vaccination prophylactique, vaccination suppressive, quarantaine...) sur le réseau, et le test d'actions ciblées sur certains élevages, sélectionnés selon leurs caractéristiques dans le réseau. Par exemple, les élevages centralisant les flux commerciaux peuvent être ciblés plus spécifiquement par des mesures de lutte. Ces simulations permettent ainsi de qualifier l'efficacité de ces méthodes en termes de nombre d'élevages finalement infectés, et de vitesse de propagation, approches que nous développerons au cours des chapitres 4 et 5.

La représentation par approche réseau permet :
- de caractériser le système et ses acteurs, ce que nous détaillerons partie 1.2.3 ;
- d'enrichir les modèles de propagation de maladies, ce que nous détaillerons partie 1.3.

1.1.3 Modèles réseaux statiques et dynamiques

Dans cette partie, nous présentons les différentes représentations que nous englobons dans l'appellation approches réseaux. Pour chacune, nous précisons sa mise en place dans le cadre de la description des échanges d'animaux d'élevages.

Chaque représentation donne accès à différents outils de description des données (partie 1.2.3), et peut également être utilisée dans le cadre de simulations de diffusions (partie 1.3).

• **Comment modéliser les interactions ?**
Le modèle le plus commun pour représenter des relations entre individus, de quelque nature que ce soit, est le modèle de graphe (que nous qualifions également de réseau statique). Il tire ses origines de la théorie des graphes, dont on attribue souvent les prémices au 18^{eme}

siècle au problème des sept ponts de Königsberg. Depuis lors, une profusion de mesures pour étudier la structure des relations représentées s'est développée.

Un graphe G est un ensemble de nœuds V, les entités étudiées, et un ensemble de liens E, leurs relations.

$$G = (V, E)$$

On notera dans la suite du manuscrit $n = |V|$ le nombre de nœuds et $m = |E|$ le nombre de liens. Un lien dans un réseau statique, que l'on dénommera lien statique par la suite, est défini comme un couple de nœuds interagissant :

$$(u, v) \in E,\ tq\ u \in V, v \in V \tag{1.1}$$

Les liens représentent des relations, dont la nature est fixée par l'utilisateur. Dans le contexte épidémiologique, un lien peut par exemple représenter une proximité ou un contact physique. C'est le cas par exemple de [Fournet and Barrat, 2014] et [Génois et al., 2015], étudiant la diffusion de maladies dans un établissement scolaire, de [Vestergaard et al., 2016, Starnini et al., 2013] en conférence scientifique, ou encore de [Liljeros et al., 2001], qui étudient la diffusion de maladies sexuellement transmissibles. Un autre exemple de relation représentée par un lien est la vente d'un animal [Christley et al., 2005, Bajardi et al., 2011, Dutta et al., 2014, **?**]. Dans la première liste d'exemples, le lien est réciproque : chaque nœud peut infecter son voisin (*i.e.* son partenaire d'interaction) au cours de ce contact. Au contraire, dans le deuxième exemple, seul le nœud vendeur peut contaminer l'acheteur, si l'animal échangé est malade. Pour prendre en compte cette non réciprocité, les liens peuvent être orientés : on distingue l'interaction du nœud u vers le nœud v, de l'interaction de v vers u. On parle alors de graphe orienté (voir le graphe de la figure 1.2).

De plus, il est possible d'affecter à chaque lien (u, v) un poids $w(u, v)$, défini comme le nombre de fois que la paire de nœuds a interagi. Le poids est appelé activité des liens statiques dans cette thèse, en référence aux mesures d'activité des nœuds, comme nous le verrons par la suite (chapitre 2). On parle alors de graphe pondéré.

Dans le contexte de l'étude des échanges de bovins entre exploitations, les nœuds sont les exploitations et les liens les échanges d'animaux. Ainsi, un nœud est lié à un autre s'ils échangent au moins un animal. Les échanges commerciaux étant orientés, les graphes orientés sont préférés pour leur représentation. En effet, cette orientation sera de grande importance dans la modélisation ultérieure de propagation de maladies.

- **Comment inclure la temporalité dans la représentation ?**

Les réseaux d'élevages étant dynamiques, les interactions n'ont pas forcément lieu d'année en année, de mois en mois, etc. entre les mêmes nœuds [Dutta et al., 2014]. D'une année sur l'autre, les nœuds ne sont donc pas reliés aux mêmes voisins. Différents modèles existent

pour prendre en compte l'évolution du réseau au cours du temps. Le plus fréquemment utilisé est la séquence de graphes (figure 1.2, centre), également appelé séquence de *snapshots* (un *snapshot* désignant un graphe de la séquence) . Il consiste à découper la période de temps étudiée en fenêtres temporelles de durée égale. Les échanges inclus dans une même fenêtre seront utilisés pour construire un graphe statique (orienté, pondéré ou non), correspondant donc à la période de temps de la fenêtre. Par exemple, il est courant d'étudier une séquence de réseaux mensuels dans le contexte de la mobilité des animaux d'élevages, comme dans [Bajardi et al., 2011], ce qui veut dire que les données sont découpées en mois, puis chaque mois est transformé en un graphe. Une séquence S peut donc se définir comme suit, avec k le nombre de réseaux de la séquence :

$$S = G_i,\ i \in \{0, 1, \ldots, k-1\}$$
$$G_i = (V_i, E_i)$$

L'atout majeur de ce modèle est qu'il permet d'utiliser les mesures de la théorie des graphes pour étudier la structure de chaque réseau statique composant la séquence. L'évolution de la structure des interactions est suivie réseau par réseau (*snapshot* par *snapshot*).

Cependant, les dates des interactions ne sont pas prises en compte au sein d'un réseau de la séquence : si l'on sait que toutes les interactions représentées ont eu lieu durant la période de temps utilisée pour construire le snapshot, on ne connaît pas leur date précise d'interaction, et donc on perd l'information sur leur ordre.

• **Modèles réseaux intégrant entièrement l'information temporelle**
Connaître l'ordre des interactions est essentiel pour restituer la notion de causalité entre les interactions et ainsi modéliser la propagation de maladies. Des modèles temporels ont donc été développés pour représenter la chronologie des interactions (voir la revue [Holme and Saramäki, 2012]). Parmi eux, on peut citer le graphe temporel de [Kostakos, 2009] : tout nœud $u \in V$ de ce réseau est représenté par un ensemble de nœuds temporels $(t, u) \in W$, un pour chaque instant $t \in T$ où l'entité a interagi. Par exemple dans la figure 1.1, le nœud A est dupliqué en deux nœuds temporels, A1 et A3, car A interagit aux temps 1 et 3. Deux types de liens coexistent dans cette représentation : les liens représentant les interactions (en noir dans la figure 1.1 à gauche), et les liens temporels (en orange dans la figure 1.1, gauche). Dans l'exemple précédant, un lien temporel relie A1 à A3. La figure 1.1, à droite, représente le graphe temporel obtenu à partir des données à gauche de la figure. Ce modèle présente l'avantage de pouvoir être étudié comme un réseau statique. Cependant, il traite de la même façon les deux types de liens, alors qu'ils représentent une information différente. Aussi, c'est souvent aux modèles de réseaux temporels s'affranchissant de la vision statique des interactions que font appel les études de la littérature.

Les *Time-Varying Graphs* (réseaux évoluant au cours du temps), notés TVG, [Casteigts et al., 2012,Santoro et al., 2011] spécifient l'occurrence dans le temps de chaque interaction.

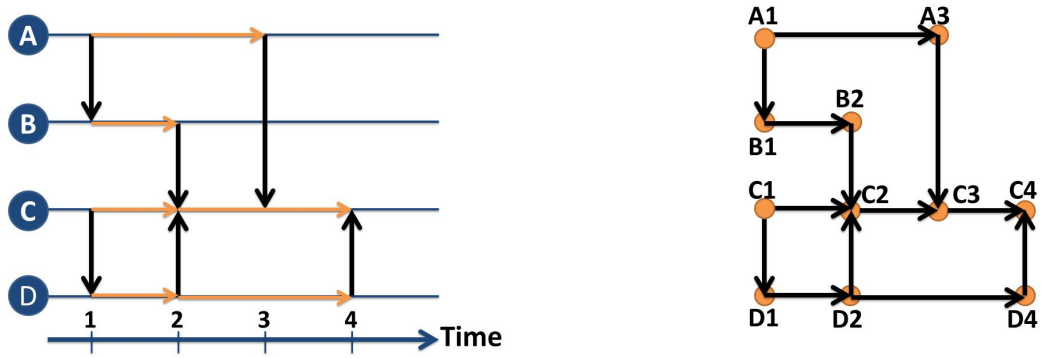


FIGURE 1.1 – Exemple de construction d'un graphe temporel, où $V = \{A,B,C,D\}$, $W = \{A1, A3, B1, B2, C1, C2, C3, C4, D1, D2, D4\}$, et $T = [1, 4]$. À gauche, on part de données décrivant des interactions : on lit par exemple que A interagit avec B au temps 1 et avec C au temps 3. En orange figurent les futurs liens temporels du graphe temporel. À droite, chaque nœud a été dupliqué aux instants de temps où il interagit : par exemple, A a été dupliqué en A1 et A3. Les liens temporels et ceux représentant les interactions sont considérés de la même façon dans le graphe temporel ainsi obtenu.

Deux fonctions permettent respectivement de renseigner les périodes d'activité des liens du graphe statique (fonction de présence des liens), et le temps nécessaire pour transmettre l'information via chaque lien (fonction de latence des liens). De même, le modèle peut inclure des fonctions de présence et de latence des nœuds, même si celles-ci sont souvent omises.

[Latapy et al., 2017] propose de représenter un réseau temporel sous la forme d'un *stream graph* :

$$S = (T, V, W, E),\ avec\ T = [t_\alpha, t_\omega],\ W \subseteq T \times V,\ et\ E \subseteq T \times V \times V$$

où W représente l'ensemble des nœuds temporels et permet donc de retrouver les instants d'activité des nœuds et ainsi les périodes où ils disparaissent du *stream graph*. Tout comme pour les TVG, le modèle peut être simplifié en supposant que les nœuds restent présents tout au long de la période de temps étudiée. On obtient alors un flot de liens (*link stream*) :

$$L = (T, V, E),\ avec\ T = [t_\alpha, t_\omega]\ et\ E \subseteq T \times V \times V$$

Un flot de liens consiste donc essentiellement en une suite d'interactions datées, autrement dit, pour un flot de liens orientés, une liste de triplets date d'interaction, nœud sortant, nœud entrant. On dénomme lien temporel un triplet $(t, u, v) \in E$. La représentation en flot de liens capture la structure mais également la dynamique des interactions, en représentant explicitement l'axe temporel (figure 1.2, droite). Ce formalisme ne suppose pas la défini-

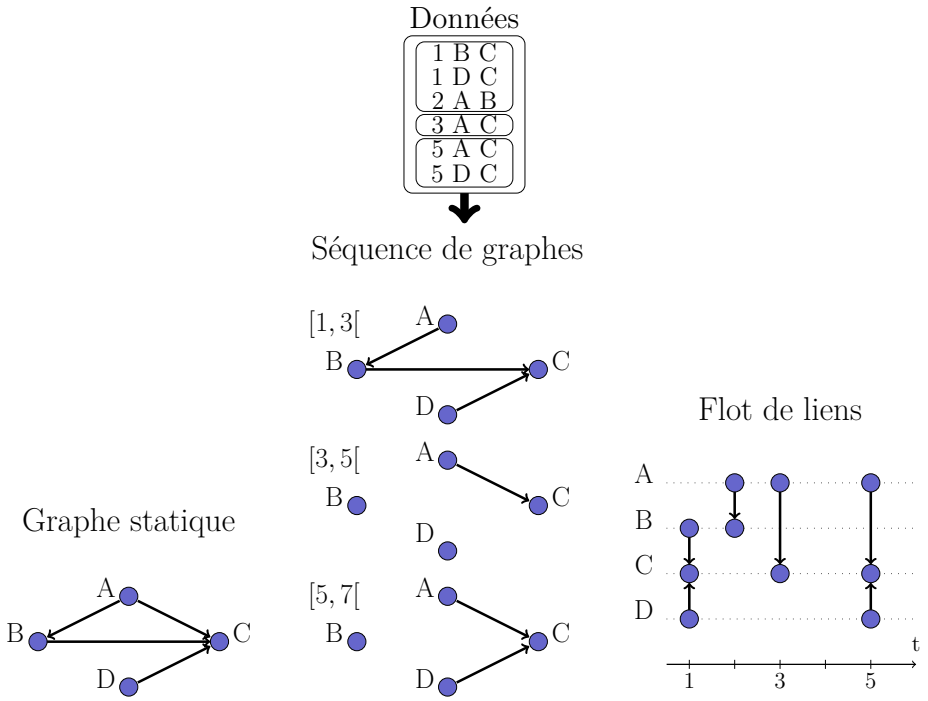


FIGURE 1.2 – Exemple de données d'échanges entre cinq nœuds, transformées selon trois modèles réseaux différents : à gauche le graphe statique correspondant, au milieu la séquence de graphes pour une durée de *snapshot* de deux unités de temps, et à droite le flot de liens. Avec le graphe à gauche, toute l'information temporelle est perdue lors de la transformation. La séquence, au milieu, conserve une part de l'information temporelle : par exemple, on lit que D a vendu un animal à C entre les temps $[1, 3[$ et $[5, 7[$. Le flot de lien à droite restitue fidèlement toute l'information temporelle : on lit que D et C ont interagi aux temps 1 et 5.

tion d'une échelle de temps caractéristique, contrairement aux séquences de *snapshots*. Nous le choisissons donc comme représentation pour cette thèse. Les échanges d'animaux sont ainsi représentés sous la forme de triplets date du mouvement, exploitation d'origine, exploitation de destination.

Si les modèles de réseaux temporels restituent fidèlement l'information temporelle contenue dans les données, ils rendent impossible l'utilisation des mesures de la théorie des graphes. Celles-ci doivent être adaptées pour prendre en compte la dimension temporelle. La question de cette adaptation est encore le sujet de nombreuses recherches.

Modèle réseau	Avantages	Inconvénients
Graphe statique	Usage possible de toutes les mesures de théorie des graphes	Perte de l'information temporelle
Séquence de graphes	Usage possible des mesures de théorie des graphes	Information temporelle incomplète. Suppose échelle de temps caractéristique
Flot de liens	Etude de la structure ainsi que de la temporalité des interactions	Mesures à adapter au contexte temporel

1.2 Modélisation de propagations

Comme souligné précédemment, la modélisation des phénomènes de propagation constitue un domaine large de la littérature. Dès que l'on se tourne vers l'étude d'un phénomène de propagation, de quelque nature que ce soit, on bénéficie ainsi de la littérature des réseaux complexes sur ces questions.

Les modèles développés sont basés sur des principes généraux qui dépassent souvent le cadre spécifique du domaine d'application. Par exemple, la propagation d'une rumeur peut être simulée suivant les mêmes règles qu'un modèle de propagation d'épidémies [Moreno et al., 2004].

Dans cette partie, nous suivons l'évolution des études portant sur la modélisation de phénomènes de diffusion, aboutissant à l'utilisation des réseaux temporels. Leur but commun est d'évaluer quelle fraction des nœuds sera atteinte par le phénomène diffusif étudié, soit par simulations (ce que l'on ferra sur la BDNI au chapitre 3), soit de manière analytique.

1.2.1 Les modèles à compartiments

Avant d'aborder ce qu'apporte l'analyse de la temporalité et de la structure des interactions, on définit le type de modèle de propagation que l'on va utiliser : les modèles à compartiments, ou à états [Keeling and Eames, 2005]. A noter que la représentation réseau sélectionnée est indépendante du modèle de propagation choisi.

L'approche la plus simple consiste à définir deux compartiments pour les nœuds, susceptible ou infecté, qui ont donné leur nom à ce modèle, susceptible-infecté (SI). Le terme infecté ne se réfère pas seulement à une maladie, mais à tout phénomène pouvant se propager. Par exemple, un nœud infecté peut correspondre à un nœud porteur d'une information [Moreno et al., 2004]. Les nœuds du compartiment susceptible ont une certaine probabilité, définie par le modèle, d'être contaminés lorsqu'ils interagissent avec des nœuds infectés. Une fois infectés, les nœuds le demeurent jusqu'à la fin de la simulation.

Lorsque les nœuds retournent à un état susceptible, un temps donné après avoir été infectés, on parle de modèle SIS. Ce type de modèle est par exemple adapté pour simuler la propagation de maladies sexuellement transmissibles [Keeling and Eames, 2005]. Dans certains cas d'étude, on peut souhaiter que les nœuds puissent sortir du compartiment infecté sans retourner à l'état susceptible. Par exemple, si [Onnela et al., 2007] utilisent un modèle SI pour simuler la propagation d'une information, [Peruani and Tabourier, 2011, Miritello et al., 2011] définissent un temps pendant lequel un nœud pourra propager l'information, avant de devenir définitivement indifférent à celle-ci. Ainsi, le modèle inclut un compartiment R, pour résistant (*recovered*) ou retiré *(removed)*, qui contient les nœuds infectés qui deviennent résistants (R) au phénomène diffusif après un temps donné. Par exemple, dans le cas des maladies, ce compartiment correspond aux nœuds guéris, devenant immunisés. [Keeling and Eames, 2005] indiquent notamment que les modèles SIR permettent de modéliser la propagation de maladies conférant une immunité à long terme, comme la rougeole. Lorsque les nœuds retournent à un état susceptible, un temps donné après avoir été guéris, on parle de modèle SIRS. Pour étudier certains cas spécifiques, comme la propagation d'une maladie donnée, d'autres compartiments peuvent être ajoutés, comme nous le verrons dans la partie 1.3.

Si l'on souhaite décrire plus en détails les mécanismes de contamination, on peut avoir recours à une description élaborée, via par exemple un système multi-agents, comme dans [Eubank et al., 2004]. Cependant, nous verrons que les objectifs fixés dans le cadre de cette thèse ne nécessitent pas l'utilisation de ce type de modélisation.

1.2.2 Description de la diffusion : des aspects structurels aux aspects temporels

La définition de modèles de propagation de maladies ne saurait être complète sans la description des possibilités de contact entre les nœuds infectés et sains. L'hypothèse la plus simple consiste à considérer que la population est *fully mixed* (c'est-à dire mélangée de manière homogène). Toutefois, il a été mesuré sur des populations réelles que tous les individus ne sont pas identiques dans leur comportement de rencontre : certains individus vont interagir avec un nombre élevés de nœuds, alors que d'autres auront un nombre faible d'interactions. Cette propriété a été très souvent décrite pour les échanges d'animaux [Rautureau, 2012, Dutta et al., 2014], et nous y reviendrons dans le chapitre suivant.

La validité de l'hypothèse que les populations sont *fully mixed* a été étudiée par la recherche des caractéristiques qui impactent la diffusion : ces caractéristiques structurelles sont-elles partagées ou au contraire fortement dépendantes des nœuds, ce qui justifierait une description plus précise des contacts afin de modéliser la diffusion ?

Tout d'abord, a été posée la question de l'impact des caractéristiques structurelles des interactions. Par exemple, [Morris and Kretzschmar, 1997] étudient l'impact du nombre de partenaires sexuels sur la transmission du VIH au sein d'une population humaine. Les auteurs montrent notamment que la vitesse de propagation augmente de manière exponentielle lorsque diminue la proportion des nœuds monogames. [Onnela et al., 2007] étudient quant à eux l'impact du poids des liens sur la diffusion. Pour ce faire, les auteurs comparent la fraction des nœuds atteinte par un modèle SI lorsque le poids de chaque lien est pris en compte, ou est remplacé par la valeur moyenne du poids. Ils montrent que l'utilisation de la valeur moyenne des poids pour toutes les interactions accélère la diffusion. Cette approximation quant au comportement d'interaction entre différentes paires de nœuds conduit donc à une surestimation des vitesses de propagations. Ces différents exemples montrent l'importance de la structure sur la dynamique de propagation.

Des études se sont également intéressées à la question de l'impact de la temporalité des interactions sur la diffusion. Ainsi dans [Vazquez et al., 2007], les auteurs étudient l'impact des distributions des temps inter-contact sur la diffusion. Celui-ci est défini comme le temps séparant deux interactions consécutives d'un même nœud. Comme nous utiliserons cette notion par la suite, nous en donnons ici la définition formelle. Soit un flot de liens $L = (T, V, E)$. Pour tout $u \in V$, on définit l'ensemble des temps de contact de u : $T_u = \{t,\ tq\ (t, u, v) \in E\} \cup \{t,\ tq\ (t, v, u) \in E\}$. Deux temps $t_i, t_{i+1} \in T_u$ sont consécutifs si $\nexists\ t' \in T_u\ tq\ t_i \geq t' \geq t_{i+1}$. On définit le i-ème temps inter-contact de u :

$$\tau_i^u = t_{i+1} - t_i \tag{1.2}$$

Les auteurs de [Vazquez et al., 2007] montrent que la diffusion suivant un modèle SI est grandement ralentie du fait l'existence d'une queue lourde dans la distribution (*fat tail distribution*) des temps inter-contact. De même, [Lambiotte et al., 2013] montrent que la vitesse des propagations d'un modèle SIR dépend de la distribution de ces temps inter-contact, et qu'elle n'est pas uniquement influencée par les valeurs de la queue épaisse. En effet, on peut considérer que ces temps représentent la probabilité qu'un lien participe à la diffusion, avant la guérison des nœuds impliqués dans cette interaction. Ces études montrent l'importance que peut prendre les caractéristiques temporelles des interactions sur la diffusion. Leur prise en compte permet ainsi de mieux comprendre les processus ayant une influence sur les propagations.

L'observation des impacts des propriétés structurelles et temporelles sur la diffusion a conduit à se poser la question de leur implication conjointe dans les processus de propagation. Par exemple, [Peruani and Tabourier, 2011] s'interrogent sur le rôle des corrélations temporelles entre les liens sur la vitesse de diffusion d'informations via des échanges téléphoniques. Dans cet exemple, les auteurs observent la probabilité plus élevée qu'un nœud en appelle un autre lorsqu'il a lui-même reçu un appel peu avant.Dans [Karsai et al., 2011],

les auteurs proposent différents modèles de référence (*null models*) pour parvenir à distinguer les effets des hétérogénéités structurelles et temporelles sur la diffusion. Un *null model* permet de rendre aléatoire certaines propriétés du réseau, ciblées spécifiquement, afin d'en étudier l'impact sur un processus de diffusion. Cette étude montre que les hétérogénéités décrites entraînent un ralentissement de la propagation d'une information, dans le cas d'échanges de messages électroniques (liens orientés) et dans le cas des conversations téléphoniques (liens non orientés). Dans [Miritello et al., 2011], les auteurs ont également recours à un *null model*, pour étudier l'impact de la distribution des temps inter-contact de leurs données d'appels téléphoniques sur la diffusion d'une information. Le null model proposé consiste à rendre aléatoire les étiquettes temporelles des interactions, de manière à détruire les caractéristiques temporelles du comportement des individus. Ils mesurent la taille moyenne obtenue par la simulation d'un modèle de diffusion SIR, en faisant varier le taux d'infection. Ils observent que les tailles observées sont différentes entre les données mesurées et les données traitées par le *null model*, du fait de l'existence de corrélations entre les interactions. Ils constatent que la diffusion est accélérée à petite échelle, alors qu'elle est ralentie à grande échelle.

Toutes ces études mettent en évidence la complexité de l'intrication entre les propriétés structurelles et temporelles des interactions. En permettant l'étude des contacts ayant véritablement eu lieu entre les nœuds et en préservant l'information temporelle sur les interactions, les réseaux temporels répondent au besoin de prendre en compte la structure comme la temporalité des interactions pour la simulation de phénomènes de diffusion. Nous les utiliserons donc pour modéliser les échanges d'animaux d'élevages entre exploitations.

1.2.3 Caractérisation des acteurs et des échanges via les approches réseaux

Dans cette partie, nous posons les définitions des propriétés structurelles des graphes et réseaux temporels qui vont être utilisées tout au long de cette thèse.

L'étude de ces propriétés permet d'obtenir des informations importantes sur l'organisation et la dynamique du réseau étudié, comme nous le verrons au chapitre 2. Elle permettra également l'identification des acteurs présentant le risque le plus fort d'être atteints et de propager des maladies, comme abordé au chapitre 4.

1.2.3.1 Éléments de structure

- **Les chemins et la distance**

Définir la notion de chemin permet d'étudier l'organisation des échanges entre les nœuds.

Dans un graphe $G = (V, E)$, un chemin $u_0 \rightsquigarrow u_k$ est une suite finie de liens, permettant de relier deux nœuds u_0 et u_k (en respectant l'orientation des liens dans le cas des graphes orientés), pour un k entier :

$$u_0 \rightsquigarrow u_k = (u_0, u_1),\ (u_1, u_2),\ ...,\ (u_{k-1}, u_k),\ avec\ (u_i, u_{i+1}) \in E \tag{1.3}$$

On dit que u_k est accessible depuis u_0 ou que u_k est atteint par u_0. Dans la suite de ce travail, les chemins sont orientés, sauf mention contraire. La notation ci-dessus fait donc référence aux chemins orientés. Dans le cas contraire, nous utiliserons la notation : $u_0 \leftrightsquigarrow u_k$.

A partir de la notion de chemin, on peut définir et mesurer :
- la longueur d'un chemin, qui est le nombre de liens le composant, soit $k+1$ pour la définition 1.3 ;
- un plus court chemin séparant une paire de nœuds : un chemin de longueur minimale permettant de relier cette paire ;
- la distance d est la longueur d'un plus court chemin (aussi appelée distance géodésique) ;

Décrire la diversité des valeurs prises par ces notions pour toutes les paires de nœuds participe à la compréhension de la structure du graphe considéré. Par ailleurs, ces notions servent à construire des mesures de l'importance des nœuds et des liens dans le graphe, comme nous le verrons par la suite.

- **Les composantes connexes**

Les composantes connexes sont des ensembles de nœuds maximaux tels qu'il existe un chemin qui relie chaque paire de nœuds de la composante. Lorsque l'on ne prend en compte que les chemins orientés, on parle de composante fortement connexe (*strongly connected component*, notée SCC). Dans un graphe orienté où l'orientation des liens n'est pas prise en compte dans la recherche des chemins, on parle de composante faiblement connexe (*weakly connected component*, notée WCC). La taille de la plus grande composante connexe (en terme de nombre de nœuds) est souvent évaluée lors des études épidémiologiques, comme nous le verrons dans le chapitre 3. Cette plus grande composante connexe étant souvent de taille très supérieure aux autres, on parle de composante géante. On note donc GSCC la plus grande SCC, et GWCC la plus grande WCC (avec G pour *Giant*). Nous reviendrons dans le chapitre 3 sur les notions de composantes connexes et comment elles permettent de décrire la structure des graphes.

- **Chemins temporels**

Le formalisme des graphes ne permet pas de prendre en compte l'ordre chronologique des liens, contrairement aux réseaux temporels. Dans ce cas dynamique, un chemin temporel $(t_0, u_0) \rightsquigarrow (t_{k-1}, u_k)$ devient une succession de liens temporels permettant de relier une

paire de nœuds (u_0, u_k), au départ d'un temps t_0 :

$$\begin{aligned}&(t_0, u_0) \rightsquigarrow (t_{k-1}, u_k) = (t_0, u_0, u_1),\ (t_1, u_1, u_2),\ ...,\ (t_{k-1}, u_{k-1}, u_k)\\ &avec\ t_0 < t_1 < \cdots < t_{k-1}\ et\ (t_i, u_i, u_{i+1}) \in E\end{aligned} \tag{1.4}$$

Ce chemin a alors une durée $t_{k-1} - t_0$, et une longueur k. Il est par nature orienté, un chemin ne pouvant remonter le temps.

On retrouve dans les réseaux temporels plusieurs notions chemin optimaux :
— Le chemin $(t_0, u_0) \rightsquigarrow (t_{k-1}, u_k))$ est le plus court (*shortest*) s'il est de longueur minimale, c'est-à-dire qu'il a le plus petit nombre de liens temporels séparant $u0$ au temps t_0 de u_k au temps t_{k-1}. Dans ce cas, on nomme distance sa longueur ;
— Le chemin le plus rapide (*fastest*), est celui de durée la plus faible ;
— On note $\mathcal{T}_{t_0}(u, (t, v))$ le temps pour atteindre (t, v) depuis u au temps t_0, défini comme suit [Latapy et al., 2017] : $\mathcal{T}_{t_0}(u, (t, v)) = t' - t_0$, où $t' \leq t$ est le plus petit temps tel qu'il existe un chemin $(t_0, u) \rightsquigarrow (t', v)$. Ce type de chemin de (t_0, u) à (t, v) est dénommé *foremost path*.

• **Cascades**

Nous définissons la notion de cascade dans un graphe, puis dans un flot de liens. Cette notion est essentielle dans la suite de ce manuscrit pour évaluer l'impact potentiel d'une diffusion sur un réseau.

Le parcours en largeur (*Breadth-first search*) est un algorithme classique qui construit à partir d'un graphe $G = (V, E)$ et d'un nœud racine $r \in V$ un arbre de plus courts chemins de r à tous les nœuds qu'il peut atteindre, comme suit. L'algorithme démarre avec une file[1] vide, à laquelle il ajoute r. Puis tant que la file n'est pas vide il prend un nœud v dans la file, et y ajoute tous ses voisins u qui n'y ont jamais été ajouté. L'ensemble des tels liens (v, u) est noté $\mathscr{C}(r)$, et il est appelé *cascade*. La taille de la cascade est le nombre de nœuds qu'elle implique, soit $|\{u, \exists(u, v)$ ou $(v, u) \in \mathscr{C}(r)\}|$.

Dans le contexte temporel, nous bornons les cascades dans le temps, c'est-à-dire que leur calcul s'arrête au bout d'un certain temps d. Nous verrons dans le chapitre 3 la motivation d'une telle définition. Nous généralisons alors le parcours en largeur à un flot de liens $L = (T, V, E)$ à partir d'une racine $r = (t, v) \in T \times V$ comme suit. L'algorithme démarre avec un ensemble vide, auquel il ajoute v, et avec un temps courant c qu'il initialise à t. Il recherche ensuite le plus petit t_x avec $c \leq t_x \leq t + d$, tel qu'il existe un lien $(t_x, x, y) \in E$ pour lequel x est dans l'ensemble et y n'y est pas ; il ajoute alors y à l'ensemble et change la valeur de c en t_x. S'il ne trouve pas de tel t_x alors l'algorithme s'arrête. L'ensemble des tels liens (t_x, x, y) est noté $\mathscr{C}_d(r)$, et il est appelé *cascade*. La figure 1.3 illustre cette notion.

1. Une file est une ensemble dans lequel les éléments sont pris dans le même ordre qu'ils y ont été ajouté, appelé aussi FIFO pour *First In first Out*.

La taille de la cascade est le nombre de nœuds qu'elle implique, soit :

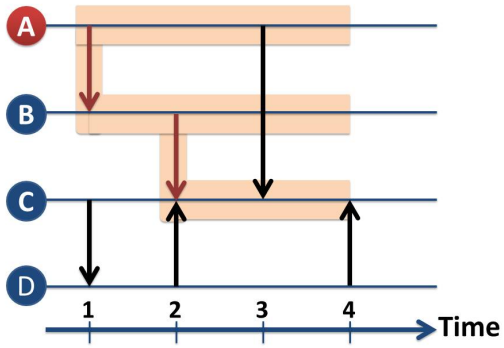


FIGURE 1.3 – Exemple de cascade au départ de A au temps 1, et d'une durée de trois unités de temps : B est atteint au temps 1, et C au temps 2. Il y a donc 3 nœuds dans la cascade.

$$|\{u, \exists(t,u,v) \text{ ou } (t,v,u) \in \mathscr{C}_d(r)\}| \tag{1.5}$$

On retrouve le terme *chaîne d'infection sortante* dans la littérature, par exemple dans [Nöremark et al., 2011], pour désigner la taille d'une cascade dans le contexte temporel.

1.2.3.2 Les centralités

L'importance des nœuds (et plus rarement des liens) est quantifiée dans un graphe au moyen de diverses notions de centralité, chacune reflétant une acception de l'importance. Dans cette partie, nous présenterons des centralités souvent utilisées pour étudier les phénomènes de diffusion.

Les mesures de centralité ont un double emploi dans l'étude des réseaux. D'une part, elles font partie des statistiques des nœuds qui sont presque systématiquement étudiées pour décrire leur comportement, par exemple dans [Rautureau, 2012, Dutta et al., 2014]. Nous y reviendrons chapitre 2. Par ailleurs, ce sont des méthodes très utilisées pour identifier les nœuds à surveiller ou à traiter en cas de crise sanitaire (cf chapitre 4).

- **Degré et activité**

Le degré permet d'identifier les nœuds ayant le plus grand nombre de voisins dans le réseau. Ces nœuds ont donc plus d'occasions de transmettre une maladie, une information etc.

En contexte dirigé, il est possible de distinguer les degrés entrant (nombre de liens arrivant au nœud) et sortant (nombre de liens partant du nœud) (figure 1.4). Le degré

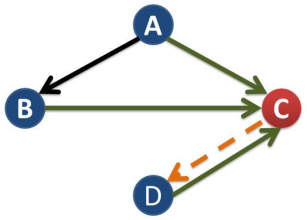


FIGURE 1.4 – Exemple de calcul de centralités sur un graphe : C a un degré total de 3, un degré sortant de 1, et un degré entrant de 3. C se trouve sur 3 plus courts chemins (de A à D, de A à C et de C à D), qui sont les uniques plus courts chemins entre ces paires de nœuds.

total est la somme des degrés entrant et sortant, et est donc égal au degré dans les graphes non orientés. Nous fixons les notations suivantes :

$$\begin{aligned} G = (V, E)\ \textit{non orienté},\ & k(u) = |\{u, (u, v) \in E\}| \\ G = (V, E)\ \textit{orienté},\ & k_{in}(u) = |\{u, (v, u) \in E\}| \\ & k_{out}(u) = |\{u, (u, v) \in E\}| \\ & k_{tot}(u) = k_{in}(u) + k_{out}(u) \end{aligned} \tag{1.6}$$

Une manière naturelle d'intégrer la dimension temporelle au degré est d'utiliser les mesures d'activité sur une période de temps donnée. Elles consistent à quantifier combien de fois un élément apparaît dans les données (figure 1.5), dans la période de temps fixée. Si l'on utilise le terme activité (noté $\mathcal{A}$) pour les nœuds, le terme poids (noté w) l'est plus souvent dans le cas des liens. Formellement, pour un flot de liens $L = (T, V, E)$:

$$\begin{aligned} \mathcal{A}_{in}(u) &= |\{t, (t, v, u) \in E\}| \\ \mathcal{A}_{out}(u) &= |\{t, (t, u, v) \in E\}| \\ w(u, v) &= |\{t, (t, u, v) \in E\}| \end{aligned} \tag{1.7}$$

- **La centralité de proximité**

Le degré et l'activité prennent en compte le nombre d'interactions avec les nœuds voisins, non le positionnement du nœud dans l'ensemble du réseau.

La centralité de proximité (*closeness centrality*, notée CC) a été définie pour détecter

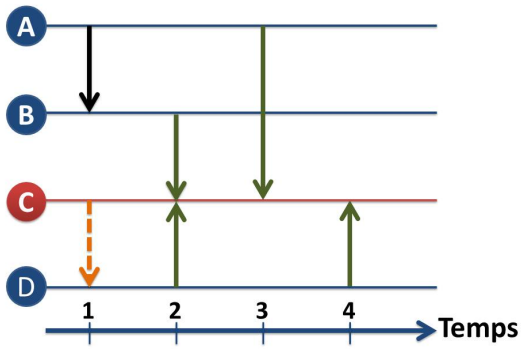


FIGURE 1.5 – Exemple de calcul de centralité sur un flot de liens : C a une activité totale de 5, une activité sortante de 1, et une activité entrante de 4. Tous les liens ont un poids de 1, sauf (lien de D à C), qui apparaît 2 fois.

les nœuds qui sont les plus proches des autres en moyenne, et pouvant ainsi les atteindre en un faible nombre de liens (figure 1.4). Elle consiste, pour un nœud, en l'inverse de la somme des distances géodésiques le séparant de tous les autres nœuds du réseau [Bavelas, 1950].

$$CC(u) = \frac{1}{\sum_{u \neq v \in V} d(u,v)}$$

La valeur de cette centralité est comprise entre 0 et 1. Plus elle est proche de 1, plus la distance moyenne entre le nœud étudié et les autres est courte, et donc plus ce nœud peut atteindre rapidement les autres, par un nombre de sauts nécessaires faible.

- **La centralité d'intermédiarité, capacité à relayer l'information**

La centralité d'intermédiarité (abrégée BC pour *betweenness centrality*) mesure l'impor-

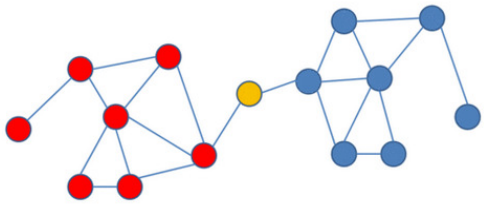


FIGURE 1.6 – Nœud (en jaune) servant de pont entre deux parties d'un graphe.

tance des nœuds et des liens selon leur capacité à être relais de la propagation. Par exemple, un nœud interagissant avec deux parties très connectées du réseau et constituant ainsi un *pont* entre elles sera détecté comme important pour la propagation, car indispensable à la transmission d'une partie à l'autre (figure 1.6). La BC est définie comme la fraction des

plus courts chemins passant par le nœud considéré [Freeman, 1977] :

$$BC(u) = \sum_{w,v \in V,\ v \neq w} \frac{\sigma_{w,v}(u)}{\sigma_{w,v}} \tag{1.8}$$

avec $\sigma_{w,v}$ le nombre de plus courts chemins entre w et v, et $\sigma_{w,v}(u)$ le nombre de plus courts chemins entre w et v passant par u. Cette centralité n'est pas nécessairement normalisée, contrairement à la centralité de proximité.

De la même manière, on peut également définir la centralité d'intermédiarité des liens : un lien est considéré comme important s'il se trouve sur une fraction élevée de plus courts chemins. Ses résultats sont souvent proches de ceux de la centralité d'intermédiarité des nœuds, ce qui peut expliquer son utilisation moins répandue pour modéliser la lutte contre la propagation des maladies.

• **Définir des centralités de proximité et d'intermédiarité temporelles**
Nous avons vu que les définitions de centralités de proximité et d'intermédiarité reposent sur la notion de plus courts chemins. Nous avons vu trois façons de généraliser le concept de plus court chemin au contexte temporel (*shortest, fastest, foremost paths*). Ces différentes définitions peuvent également être combinées : par exemple, [Latapy et al., 2017] combine les notions de chemins les plus courts et plus rapides (shortest fastest paths) pour définir la BC dans les flots de liens.

Une manière simple de prendre en compte l'évolution d'un graphe au cours du temps est de considérer une séquence de graphes correspondante, et de calculer sur chaque *snapshot* les centralités. C'est l'alternative que nous choisirons dans le chapitre 4. Le temps de calcul des plus courts chemins pouvant s'avérer coûteux, [Lee et al., 2016] proposent de mettre à jour les valeurs de BC à chaque *snapshot*. Pour ce faire, il faut calculer au préalable tous les plus courts chemins et les conserver en mémoire, ce qui aboutit finalement à une forte complexité mémoire et limite l'utilisation de cette méthode sur de grands jeux de données.

Au lieu d'utiliser les notions statiques de la CC et de la BC sur une séquence de *snapshots*, [Tang et al., 2010] proposent des définitions basées sur la notion de plus courts chemins temporels, calculés sur plusieurs *snapshots* successifs pour évaluer l'importance du nœuds au cours de la période considérée : la notion de chemins les plus rapides (*fastest*) est utilisée pour calculer la CC et la BC. Cependant, ces définitions temporelles sont très coûteuses en temps de calcul.

[Santoro et al., 2011] généralisent la BC et la CC aux *Time Varying Graphs* : comme pour leurs équivalents statiques, elles représentent respectivement la fraction des plus courts chemins temporels passant par un nœud, et la distance temporelle aux autres nœuds. Les auteurs laissent libre le choix de la notion de plus court chemin temporel retenu (court,

rapide, *foresmost*), et donc de distance temporelle (distance, durée, ou $\mathcal{T}$). De la même façon, [Latapy et al., 2017] définissent pour les flots de liens la CC et la BC en reprenant la définition statique : la CC évalue la distance des *shortest foremost paths* reliant le nœud étudié aux autres nœuds, et la BC devient la fraction des *shortest fastest paths* passant par le nœud étudié.

À ce jour, il n'existe pas de consensus quant aux définitions temporelles à retenir. De plus, les algorithmes correspondants ont une forte complexité en temps et en espace.

• **La centralité de vecteurs propres, de Katz, et *PageRank***
Alors que le degré prend en compte dans son calcul toutes les interactions de la même façon, on peut se demander si cela est justifié sachant que tous les nœuds n'ont pas la même importance, de par des comportements d'interaction différents.

La centralité de *Katz* a été développée [Katz, 1953] sur l'idée qu'un nœud est central s'il permet d'accéder en peu de sauts à un grand nombre de nœuds. A partir du nœud étudié, on mesure la longueur des chemins le reliant aux autres nœuds du réseau. Contrairement aux centralités de proximité et d'intermédiarité, tous les chemins sont ainsi pris en compte dans le calcul. Un facteur d'atténuation permet de donner plus d'importance aux chemins de faible longueur dans le calcul de la centralité. Selon cette notion, un nœud central ayant un grand nombre de liens sortants transmettra son score élevé à tous les nœuds atteints par ces liens. Cette transmission à tous les nœuds de destination sans condition peut poser question, comme cela a été le cas dans le contexte de l'évaluation de l'importance des pages webs. En effet, selon l'exemple de [Newman, 2010], si une page web très centrale comme *Yahoo !* pointe vers des millions de pages, il faudrait que les scores de centralité de celles-ci soient dilués en raison de cet ensemble important de liens sortants.

D'autres notions reposant sur une définition auto-cohérente ont été proposées par la suite : un nœud est estimé important s'il est lié à d'autres nœuds importants. Par exemple, le *PageRank* [Brin and Page, 1998] définit l'importance d'un nœud selon sa capacité à recevoir des liens ayant pour origine des nœuds u_i eux-mêmes importants. Son score sera d'autant plus élevé que les nœuds u_i ont peu de liens sortants. Ainsi, les scores de centralité ne sont pas surévalués comme dans l'exemple des pages webs connectées à la page *Yahoo !* [Newman, 2010], cité précédemment.

1.2.3.3 k-core statique et temporel

Les mesures de centralités sont les plus utilisées pour étudier la propagation des maladies dans les réseaux d'élevages. Cependant, l'étude de phénomènes de propagation a vu croître l'intérêt pour un autre type de mesure d'importance des nœuds : la décomposition en k-cores [Bollobás, 1984]. Cette décomposition consiste à supprimer successivement tous

les nœuds d'un degré inférieur à k : une fois ces nœuds supprimés, le degré est recalculé sur le réseau, et si les nœuds restants obtiennent alors un de degré inférieur k, ils sont également supprimé, etc., jusqu'à ce qu'il ne reste plus qu'un ensemble C^k des nœuds de degré supérieurs ou égal à k, appelé k-core. Ainsi, on commence par supprimer de manière itérative tous les nœuds de degré 1, ces nœuds supprimés auront un indice de 1, puis tous ceux de degré 2, etc., jusqu'à ce que tous les nœuds aient été supprimés et se soient vu attribuer un indice lors de la décomposition. [Kitsak et al., 2010] montrent alors que le degré d'un nœud n'est pas le facteur le plus important pour juger de son potentiel de propagation, mais plutôt son indice de k-core, indépendamment de son degré. Plus le nœud a un indice élevé, plus son potentiel de propagation est important. De plus, [Kitsak et al., 2010] montre que le nombre de nœuds atteints augmente de manière importante lorsque la diffusion atteint un k-core jusque là épargné, alors qu'il augmente faiblement voire reste relativement stable lorsque la propagation demeure dans des k-core déjà rencontrés.

[Latapy et al., 2017] définit la notion de k-core pour les *stream graphs*. Au lieu de mesurer le degré statique défini précédemment, les auteurs calculent à la place le degré instantané, *i.e.* le nombre d'interactions du nœud considéré à un temps t. La décomposition se déroule ensuite selon le même principe.

1.3 Modélisation de la propagation de maladies animales via les approches réseaux

Dans cette partie, nous restreignons le champ d'étude au cas spécifique de la diffusion de maladies animales. Ce domaine nécessite en effet certains ajustements pour décrire de manière plus réaliste la propagation de ce type de maladies.

1.3.1 Maladies animales : exemples de modèles à compartiments

Nous avons vu en partie 1.2.1 le concept de modèle à compartiments. Lorsque l'on étudie des cas spécifiques de maladies, des compartiments supplémentaires et les règles de transitions correspondantes peuvent être rajoutées, pour plus de réalisme. On trouve ainsi les compartiments M (immunisé passif) et E (infecté en période de latence). Par exemple, [Kiss et al., 2006] modélisent la fièvre aphteuse par un modèle SEI. De même, [Brooks-Pollock et al., 2014] utilise un SEI pour modéliser la propagation de la tuberculose bovine. Dans [Harvey et al., 2007] et [Rautureau, 2012], le modèle de propagation de la fièvre aphteuse est détaillé plus encore : en plus des compartiments S, I et R, le modèle comprend le compartiment L (infecté mais non infectieux), I (infectieux avec signes subcliniques), J

(infectieux avec signes cliniques).

La difficulté de la modélisation des maladies consiste principalement à définir des taux d'infections, des temps de guérison etc. qui soient pertinents : il faut soit avoir accès à des données sur l'enregistrement des animaux infectés par une maladie, soit trouver dans la littérature des paramètres qui puissent correspondre à l'étude menée. [Brooks-Pollock et al., 2014] décrivent ainsi comment, en se basant sur leurs estimations personnelles et celles de littérature, ils ont fixé les paramètres de leur modèle SEI de la fièvre aphteuse (taux d'infection, taux de passage du stade latent à infectieux, paramètres de transmission environnementaux). [Rautureau, 2012] propose quant à elle des paramètres pour son modèle de propagation de la fièvre aphteuse. Les modèles ainsi paramétrés sont donc spécifiques au cas étudié, de même que les conclusions qui en découlent.

1.3.2 Modélisation des contacts au sein d'une exploitation

Nous avons présenté les modèles à compartiments et les différentes représentations des contacts entre les exploitations. Dans le cas des maladies animales, il existe cependant un deuxième niveau de propagation : au sein des exploitations. Dans le cas de maladies à diffusion plus lente comme la diarrhée virale bovine [Courcoul and Ezanno, 2010] ou la paratuberculose [Beaunée et al., 2015], ce niveau a son importance dans la modélisation pour en augmenter le réalisme.

Rares sont les jeux de données enregistrant les contacts entre animaux au sein d'un même élevage. La banque de données Crawdad [2] propose par exemple des données de contact entre bovins [Wietrzyk and Radenkovic, 2007]. Cependant, peu d'animaux disposent d'une balise GPS, ce qui rend difficile la modélisation précise de la dynamique de contact intra-élevage. Il est donc difficile d'utiliser une approche réseau pour décrire la transmission intra-exploitation. Une autre représentation des contacts doit donc être utilisée.

Pour l'exemple des exploitations laitières et de la diarrhée virale bovine, les modèles sont adaptés aux conditions d'élevage typiquement rencontrées sur le terrain. Ainsi, [Ezanno et al., 2007] regroupent les animaux en 5 groupes selon leur âge et leur stade de lactation. Ils utilisent 5 compartiments pour décrire leurs états au cours du temps : susceptible, infecté de manière transitoire, infecté de manière permanente, immunisé, et protégé par les anticorps maternels. Chaque groupe d'animaux a des probabilités spécifiques de passer d'un état à un autre. [Courcoul and Ezanno, 2010] reprennent ce modèle intra-troupeau et le complètent par la modélisation de la propagation par introduction d'animaux (par achat) et par contact au pâturage, afin de mieux évaluer la part respective de chaque voie

2. https ://www.crawdad.org/

de propagation dans la diffusion. Ils montrent que la maladie ne peut persister sur le long terme dans le réseau sans le concourt des échanges d'animaux, et que ceux-ci augmentent également la durée de l'infection dans la population. De même, [Beaunée et al., 2015] utilise un modèle à compartiments pour décrire la diffusion de la paratuberculose au sein des exploitations, avec 5 groupes d'âges pour les animaux et 6 compartiments : susceptible pour les animaux âgés de moins d'un an, résistant (immunisé) pour les animaux de plus d'un an, infecté de manière transitoire juste après l'infection, infecté mais non infectieux (L), infectieux sans symptômes (IS), infectés cliniquement visible (IC). La transmission inter-exploitations est quant à elle modélisée grâce à une approche réseau. Les auteurs concluent que le risque de contamination par l'introduction par achat d'animaux malades dépend de la prévalence [3] dans la population et du renouvellement des animaux dans l'exploitation : si le renouvellement est élevé et la prévalence faible, l'impact de la contamination par échange d'animaux diminue ; si la persistance est élevée, le risque est d'autant plus accru que le renouvellement des bêtes est faible.

1.4 Positionnement par rapport à la littérature

Au cours de ce chapitre, nous avons vu que de nombreux efforts ont été faits pour comprendre la façon dont les propriétés structurelles et temporelles d'un réseau impactent conjointement la diffusion sur ce-dernier.

Dans cette thèse, notre but est d'évaluer l'impact de la prise en compte de l'information temporelle dans l'étude d'un réseau d'échanges d'animaux d'élevage. Nous chercherons ainsi à comprendre comment les phénomènes de diffusion décrits sur d'autres jeux de données dynamiques vont se traduire sur la BDNI. Nous aurons pour but de comprendre, par l'étude des cascades en contexte temporel, l'entremêlement des propriétés structurelles et temporelles sur le flot de liens de la BDNI.

Pour ce faire, on ne souhaite pas ici simuler la propagation d'une maladie réaliste, mais au contraire, déterminer des caractéristiques générales de la diffusion sur le réseau, dans le contexte temporel. Ainsi, nous modéliserons la propagation par un modèle SI. Il ne nous est pas nécessaire de simuler la diffusion intra-exploitation.

Nous choisissons de nous positionner dans le pire cas de propagation, c'est-à-dire un modèle SI avec un taux d'infection de 1. Une première motivation pour justifier son usage est que mesurer le nombre de nœuds atteints avec un tel modèle est alors déterministe. Il devient un élément de description du réseau, permettant de comprendre sa structure et de tester l'accessibilité des nœuds. Nous verrons dans le chapitre 3 que ce positionnement est

3. Nombre d'infectés par une maladie à un moment donné

souvent adopté, notamment par les études mesurant la taille des composantes connexes. De plus, certaines propriétés qualitatives resteraient vraies avec la plupart des modèles plus réalistes, par exemple ceux ayant un taux d'infection basé sur des observations de maladies réelles. On s'attend notamment à ce que soit le cas des nœuds identifiés comme jouant un rôle majeur dans les diffusions. Il n'est donc pas forcément nécessaire d'atteindre un niveau de détail élevé pour étudier la vulnérabilité du réseau à la propagation, selon les objectifs poursuivis. Ce modèle SI semble donc répondre à nos objectifs.

Chapitre 2

Description et analyse des données

Sommaire

En France, le ministère de l'agriculture et de l'alimentation est chargé d'appliquer la directive européenne [EU, 2000] visant à assurer la traçabilité des changements d'exploitation des animaux d'élevage. La base de données nationale d'identification des animaux (BDNI) identifie donc tous les bovins présents sur le territoire, et enregistre chacun de leurs changements d'exploitation.

Une exploitation est par définition un lieu où des bovins sont détenus, élevés, et entretenus [Ministère de l'agriculture et de l'alimentation, 2004]. Ce terme fait donc référence à un lieu géographique, et non pas à une personne physique. Différents types d'exploitation sont répertoriés dans la BDNI [Ministère de l'agriculture et de l'alimentation, 2004] :

— les élevages sont des lieux où les animaux sont détenus en vue d'engraissement ou de reproduction ;
— les centres de rassemblement sont des lieux, distincts des élevages, où les animaux sont regroupés en vue de former des lots, destinés à la vente. Les animaux n'y séjournent pas plus de 30 jours ;
— les marchés sont des centres de rassemblement où les animaux ne restent généralement pas plus d'un jour, et où les animaux restent sous la responsabilité de la personne physique gérant leur exploitation d'origine ;
— les abattoirs sont les lieux où les animaux sont abattus ;
— les exploitations d'équarrissage sont les lieux où les carcasses sont déchargées.

L'origine et la destination des bovins déterminent le type d'échange effectué et la table dans lequel ce changement d'exploitation sera enregistré. Ce sont ces tables enregistrant les mouvements qui vont nous permettre de construire les réseaux d'échanges. Nous les présentons dans une première partie. Puis nous détaillons le pré-traitement effectué, nécessaire à l'utilisation des données. Enfin, nous décrivons les données à l'aide de mesures statistiques, présentées dans la partie 1.2.3.

2.1 Description des tables et correction des incohérences

L'extraction de la BDNI reçue dissocie chaque type d'échange d'animaux pour former les tables correspondantes. Ainsi, cinq tables enregistrent les mouvements : entre élevages, avec un centre de rassemblement, avec un marché, vers un abattoir, et vers un équarrisseur. Chacune présente une structure qui lui est propre.

2.1.1 Table des mouvements entre élevages

Un enregistrement (un n-uplet de la table) donne les informations sur une période de détention d'un animal dans un élevage. Entre autres attributs, un enregistrement correspond à un animal, le numéro de l'exploitation, sa date d'entrée sur l'exploitation, et sa date de sortie le cas échéant. Par rapport à notre objectif d'étudier le réseau des échanges entre exploitations, on voit donc que le numéro de l'élevage de destination en cas de sortie de l'animal ne figure pas explicitement. Nous verrons quel pré-traitement a été mis en place pour reconstituer l'information sur l'origine et la destination d'un échange, nécessaire à la construction du réseau.

220 163 983 enregistrements sont rassemblés dans cette table. Nous avons relevé plu-

sieurs problèmes :

— Présence d'exploitations d'autre type qu'élevage, dans cette table réservée aux échanges entre élevages ;
— Les dates de sorties ne sont pas fiables : un exploitant peut ne pas déclarer la date exacte à laquelle l'animal a été vendu, souvent par oubli ;
— Un même animal peut avoir plusieurs dates de naissance.

Pour répondre à la non fiabilité des dates de sorties, nous supposons que les dates d'entrées sont fiables et travaillons uniquement avec elles : les enregistrements sont triés par animal puis par date d'entrée, ce qui permet de reconstituer les trajets des animaux au cours de leur vie. Concernant les dates de naissance, nous conservons la plus ancienne. Avec ces vérifications, $0,07\%$ des enregistrements ont été enlevés. L'impact des incohérences est donc marginal pour l'étude des échanges entre élevages.

2.1.2 Table des mouvements impliquant un marché

Dans cette table, un enregistrement indique notamment le numéro d'identification de l'animal, le numéro du marché sur lequel il est transféré, et la date du mouvement. 13 550 508 enregistrements sont rassemblés dans cette table. À la réception au ministère des informations renseignées par les exploitants, aucune vérification n'est faite sur les champs. Il peut donc manquer des informations, y avoir des erreurs lors de la saisie, etc.

Pour reconstituer le trajet de chaque animal, c'est-à-dire depuis quelle exploitation puis vers quelle autre exploitation il transite après son passage sur un marché, nous croisons l'information de cette table avec celle des mouvements entre élevages : en triant les données par animal puis par temps croissant, nous pouvons identifier l'exploitation ayant vendu l'animal au marché, puis celle l'ayant acheté. Cette méthode fonctionne indépendamment du type des exploitations d'origine et destination.

2.1.3 Table des mouvements passant par un centre de rassemblement

Un enregistrement indique le numéro d'identification de l'animal, le numéro du centre de rassemblement, la date du mouvement, le type du mouvement (entrée ou sortie). Ainsi chaque enregistrement correspond soit à une entrée, soit à une sortie du centre. Cette table comprend 141 780 457 enregistrements. De même que pour la table des marchés, le ministère ne fait aucune vérification sur les champs entrés par les acteurs.

Nous avons pu identifier 7 types d'incohérences concernant l'enregistrement des dates d'entrée et de sortie, ce qui n'exclut nullement l'hypothèse qu'il puisse en exister d'autres.

Nous précisons le traitement effectué lorsque cela est possible, sinon, les enregistrements incohérents sont supprimés :

1. Pour le même animal, deux mouvements d'entrée se suivent dans le même centre de rassemblement. Si un mouvement de sortie suit l'entrée la plus tardive, ces deux enregistrements seront conservés normalement. La première entrée est supprimée. 214 436 occurrences (soit $0,15\%$ des enregistrements) ;
2. Pour le même animal, deux sorties se suivent depuis le même centre. La première sortie est conservée si elle suit un mouvement d'entré ne soulevant pas d'erreur. La deuxième n'est pas conservée. 1925 occurrences ($< 0,01\%$) ;
3. Enregistrement d'un mouvement de sortie pour un animal n'étant pas encore entré dans un centre. L'enregistrement n'est pas conservé. 2 204 393 occurrences ($1,6\%$) ;
4. Pour un même animal, deux entrées se succèdent, mais dans des centres différents. Une des deux entrées peut être une erreur, ou la sortie du premier centre n'a pas été notifiée. On supprime la première entrée ; pour la seconde, si elle est suivie d'une entrée qui ne soulève pas d'erreur, elle est conservée normalement. 712 509 occurrences ($0,5\%$) ;
5. Pour un même animal, deux sorties se succèdent depuis deux centres différents. L'enregistrement n'est pas conservé. 848 occurrences ($< 0,01\%$) ;
6. Pour un même animal, une sortie succède bien à une entrée, mais pas du même centre. Il y a soit une double erreur (non notification de la sortie du premier centre et non notification de l'entrée dans le deuxième centre), soit il y a eu une erreur dans la saisie du numéro du centre dans l'un des deux enregistrements, soit ces deux problèmes coexistent. Les enregistrements ne sont pas conservés. 204 840 occurrences ($0,15\%$) ;
7. Un animal entre dans un centre sans jamais sortir, de ce centre ou d'un autre. L'enregistrement n'est pas conservée. 2 559 769 occurrences ($1,8\%$) ;

Au total, $4,16\%$ des enregistrements de cette table sont concernés par ces incohérences.

2.1.4 Autres tables de mouvements

Nous nous intéressons dans le cadre de cette thèse à la propagation de maladies, et donc aux échanges en permettant la diffusion. La BDNI comporte une table pour les mouvements vers les abattoirs, et une vers les équarrisseurs. Nous ne les prendrons pas en compte dans la modélisation : ce sont des impasses pour la diffusion. Toutes les incohérences les concernant ne portent donc pas préjudice à la suite de l'analyse. Je les présente tout de même succinctement, afin de recueillir dans un même document les incohérences majeures que l'on peut trouver à ma connaissance dans la BDNI.

• **Entrées dans les abattoirs**
Un enregistrement donne le numéro de l'abattoir, le numéro national de l'animal reçu, la date entrée, la date abattage, et l'exploitation d'origine.

Incohérences constatées :
— La date d'entrée est parfois non renseignée. Il ne reste alors que la date d'abatage comme indice sur la date du mouvement ;
— Un abattoir peut acheter des bovins à lui-même ;
— Les animaux peuvent être abattus plusieurs fois à des dates différentes. Ceci pose d'autant plus problème si l'animal a fait des mouvements entre exploitations entre la première date d'abattage et la dernière date d'abattage.

Au total, seulement 1705 enregistrements présentaient des incohérences, sur les 66 639 052 contenues dans la table, soit un pourcentage d'erreur inférieur à $0,01\%$.

• **Entrées chez les équarrisseurs**
Un enregistrement indique le numéro d'identification de l'animal, l'exploitation de provenance, le numéro de l'équarrissage, et date d'enlèvement.

Incohérences constatées :
— Les animaux peuvent être équarris plusieurs fois, à des dates différentes et en provenance de différents élevages ;
— L'exploitation de provenance de l'animal et le numéro de l'équarrissage peuvent être identiques.

Sur les 17 107 146 enregistrements présents dans la table, seuls 5584 ($0,03\%$) étaient concerné par des problèmes d'incohérences.

2.2 Mise en forme des données

2.2.1 Reconstitution des échanges

Nous avons décrit précédemment comment nous avons traité les incohérences trouvées dans la BDNI. Nous sommes donc à la deuxième étape de la figure 2.1 récapitulative. Afin d'obtenir des données sous un format unique, nous fusionnons les informations extraites des tables des échanges entre élevages, marchés et centre de rassemblement. Un tri selon le numéro d'identification des animaux, puis les dates d'échange est effectué. Ensuite, nous effectuons un pré-traitement afin d'obtenir des données de la forme : date de l'échange, exploitation d'origine, exploitation de destination, numéro de l'animal. Cette mise en forme fait apparaître parfois un nouveau type d'incohérence, lorsque l'échange d'un animal d'une exploitation s'effectue vers la même exploitation. L'échange n'est alors pas conservé, c'est

le cas de $0,8\%$ d'entre eux. Les données sont ensuite triées selon la date de l'échange, afin de rétablir l'ordre chronologique des interactions.

2.2.2 Échelle de représentation

Les données sont maintenant sous la forme de quadruplets, notés q dans le tableau 2.1 : date de l'échange, exploitation d'origine, exploitation de destination, numéro de l'animal. Si deux exploitations échangent plusieurs animaux, plusieurs quadruplets se succèdent donc, chacun représentant le changement d'exploitation d'un animal. D'un point de vue réseau, cela consiste à représenter des liens temporels multiples entre deux nœuds. Nous faisons le choix de ne pas conserver les numéros d'identification des animaux et d'agréger les quadruplets de sorte d'obtenir des liens temporels orientés (des triplets date, origine, destination). Nous retombons ainsi dans le cadre usuel des modèles réseaux temporels. Cette représentation sera en effet suffisante pour répondre à nos objectifs de recherche. Un échange ne correspond donc pas au changement d'exploitation d'un unique animal, mais à celui d'un lot d'animaux, comprenant un animal ou plus.

2.2.3 Période d'analyse

Dans les données obtenues, on observe des incohérences concernant les dates des échanges. Notamment, on trouve des échanges d'animaux datant des années 80, donc antérieurs au dispositif d'enregistrement dans la BDNI. Le ministère nous ayant mis en garde sur certains problèmes de fiabilité qui ont existé après la mise en place de la BDNI, nous choisissons l'année 2005 comme point de départ, car cette année est suffisamment ultérieure à la date de mise en œuvre de la BDNI dans les années 2000. De plus, c'est l'année la plus ancienne de la BDNI à avoir été étudiée par le passé [Rautureau, 2012]. Nous étudions donc les données de 2005 à la dernière année obtenue, 2015. A partir des études de [Rautureau, 2012] (année 2005) et de [Dutta et al., 2014] (2005 à 2009), nous pouvons vérifier la cohérence des résultats obtenus par la suite avec l'extraction de la BDNI que nous avons obtenue. De 48 194 476, nous passons à 32 610 834 liens temporels (échanges de lots d'animaux) lorsque l'on restreint les données à la période de 2005 à 2015 (figure 2.1).

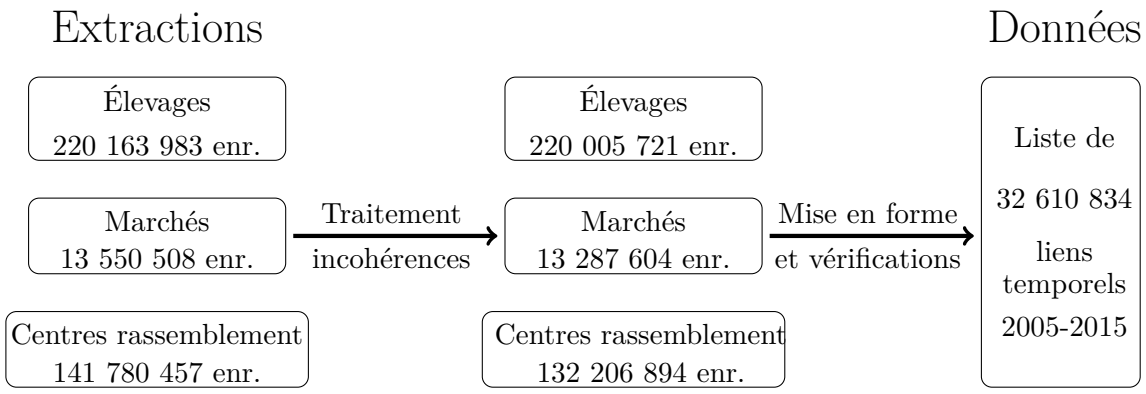


FIGURE 2.1 – Volume des données de la BDNI avant et après pré-traitement. On note *enr.* les enregistrements.

TABLE 2.1 – Statistiques des nœuds et des liens de la BDNI, des années 2005 à 2015.
Notations : n représente le nombre de nœuds actifs une année donnée, q le nombre de quadruplets *date, nœud d'origine, nœud de destination, numéro du bovin échangé*, $|E|$ le nombre de liens temporels, m le nombre de liens statiques ; la réciprocité est notée r, et le degré k ; ma pour les marchés, c pour les centres de rassemblement, et el pour les élevages. Par suite, on note n_{ma} le nombre de marchés du réseau, etc.

	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
n	245 821	236 903	227 776	218 995	211 163	205 057	198 451	192 520	186 904	182 452	176 771
n tq $k_{out} > 0$	232 900	224 424	214 775	204 924	198 205	193 402	187 101	181 613	176 545	172 154	167 193
n tq $k_{in} > 0$	143 844	138 910	132 834	127 16	121 715	117 603	115 138	112 967	109 999	107 629	101 481
q	10 398 841	10 422 425	9 277 034	9 097 858	9 064 502	9 426 612	9 820 428	9 722 380	9 883 533	10 270 761	10 298 373
$\|E\|$	3 355 680	3 212 637	2 897 540	2 739 938	2 689 191	2 746 280	2 809 489	2 725 040	2 679 914	2 719 708	2 652 131
m	1 646 510	1 512 351	1 404 920	1 292 155	1 258 145	1 239 570	1 221 175	1 123 725	1 086 816	1 060 354	1 037 465
$\frac{q}{n}$	42,30	44,00	40,73	41,54	42,93	45,97	49,49	50,50	52,88	56,30	58,26
$\frac{\|E\|}{n}$	13,65	13,56	12,72	12,51	12,74	13,39	14,16	14,16	14,34	14,91	15,00
$\frac{\|u,u\in GSCC\|}{n}$	0,44	0,44	0,43	0,41	0,41	0,41	0,41	0,44	0,44	0,41	0,42
r	0,11	0,11	0,11	0,10	0,10	0,10	0,11	0,12	0.13	0,12	0,14
$\frac{n_{ma}}{n}$	0.0003	0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004
$\frac{n_c}{n}$	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.008	0.008	0.008	0.008
$\frac{\|u_{ma},u_{ma}\in GSCC\|}{n_{ma}}$	1	0.99	0.95	0.93	0.92	0.86	0.83	0.79	0.77	0.76	0.75
$\frac{\|u_c,u_c\in GSCC\|}{n_c}$	0.73	0.71	0.69	0.66	0.63	0.63	0.65	0.67	0.71	0.73	0.74
$\frac{\|u_{el},u_{el}\in GSCC\|}{n_{el}}$	0.43	0.44	0.43	0.41	0.41	0.41	0.42	0.43	0.44	0.43	0.42

TABLE 2.2 – Fréquence des liens selon leur type.

Année	entre élevages	élevage →marché	marché →élevage	marché →centre	centre →marché	élevage →centre	centre →élevage	entre centres	entre marchés
2005	0,277	0,083	0,022	0,017	0,011	0,465	0,074	0,050	0,002
2006	0,262	0,083	0,021	0,0184	0,011	0,477	0,072	0,055	0,002
2007	0,278	0,077	0,020	0,018	0,010	0,475	0,070	0,050	0,002
2008	0,275	0,073	0,021	0,017	0,010	0,486	0,071	0,046	0,002
2009	0,273	0,072	0,020	0,017	0,009	0,490	0,069	0,048	0,002
2010	0,262	0,068	0,019	0,018	0,009	0,506	0,066	0,0519	0,002
2011	0,253	0,060	0,016	0,016	0,008	0,524	0,068	0,053	0,002
2012	0,231	0,053	0,014	0,015	0,008	0,547	0,071	0,058	0,001
2013	0,214	0,051	0,015	0,017	0,010	0,554	0,074	0,064	0,001
2014	0,199	0,046	0,013	0,018	0,010	0,564	0,077	0,071	0,001
2015	0,204	0,043	0,012	0,017	0,010	0,568	0,074	0,071	0,001

2.3 Analyse des données

Cette partie a pour but de dégager les caractéristiques principales de la BDNI pouvant avoir un impact sur la diffusion. Leur étude servira donc à éclairer les résultats obtenus dans les chapitres suivants.

Notre extraction de la BDNI nous donne accès à 11 ans de données d'échanges de bovins, de 2005 à 2015. A notre connaissance, seules deux études de la BDNI ont été menées jusqu'à ce jour, par [Rautureau, 2012] pour l'année 2005, et par [Dutta et al., 2014] pour les années 2005 à 2009. Étant donné que nous avons reçu une extraction différente de celles de ces études, nous chercherons à vérifier la concordance des mesures effectuées entre nos données et les leurs.

2.3.1 Activité des nœuds et des liens

Afin de mieux comprendre l'organisation des échanges, nous mesurons le nombre de nœuds qui interagissent au moins une fois sur une période donnée. Ces nœuds sont dit actifs sur cette période. On distingue dans la définition suivante les nœuds ayant une activité d'achat et ceux ayant une activité de vente. De même, nous mesurons le nombre de liens actifs sur une période donnée. Pour un flot de liens $L = (T, V, E)$ et un intervalle de mesure $[t_i, t_f]$, on définit ainsi :

$$\begin{aligned} O^V(t_i, t_f) &= |\{u,\ tq\ \exists t \in [t_i, t_f]\ et\ v \in V\ tq\ (t, v, u)\ ou\ (t, u, v) \in E\}| \\ O^V_{in}(t_i, t_f) &= |\{u,\ tq\ \exists t \in [t_i, t_f]\ et\ v \in V\ tq\ (t, v, u) \in E\}| \\ O^V_{out}(t_i, t_f) &= |\{u,\ tq\ \exists t \in [t_i, t_f]\ et\ v \in V\ tq\ (t, u, v) \in E\}| \\ O^E(t_i, t_f) &= |\{(u, v),\ tq\ \exists t \in [t_i, t_f]\ et\ v \in V\ tq\ (t, u, v) \in E\}| \end{aligned} \tag{2.1}$$

Dans cette partie, on divise chaque année en intervalles de temps de durée variable (d'un jour à un mois), et pour chaque intervalle, on mesure O^E et O^V.

Comme souligné par de nombreuses études de données d'échanges d'animaux [Christley et al., 2005,Robinson et al., 2007], la BDNI présente elle aussi des périodes de forte activité succédant à des périodes de faible activité :

— Lorsque O^V est calculé sur des intervalles d'un mois, l'activité est plus forte en automne et au début du printemps. Elle est plus faible en été. On retrouve des résultats cohérents avec [Dutta et al., 2014], qui indique que mars présente l'activité la plus forte, et août la plus faible, ainsi qu'avec [Robinson et al., 2007], qui observe des pics d'activité au printemps et à l'automne ;

— Pour des intervalles de calcul d'une semaine (figure 2.2, à droite), la dernière se-

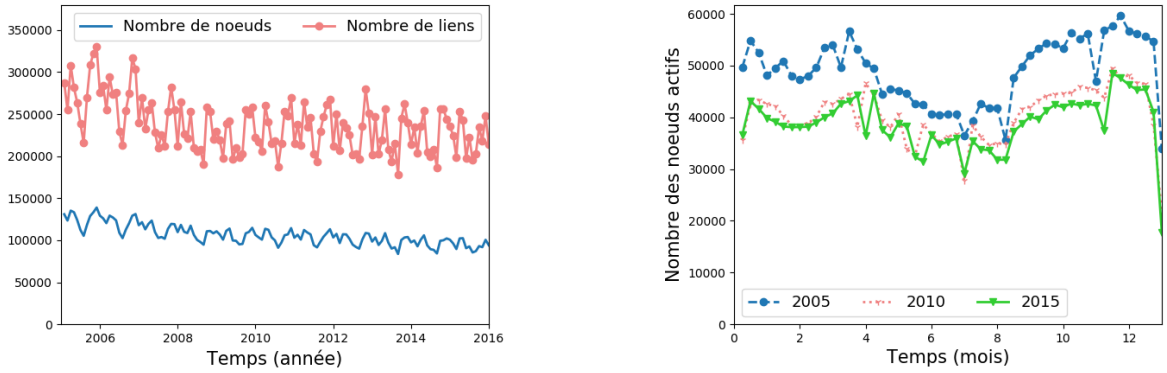


FIGURE 2.2 – Nombres de nœuds et de liens actifs mesurés tous les mois (à gauche) et toutes les semaines (à droite).

maine de décembre, la première semaine de juillet et les deux premières semaines d'août sont les périodes de plus faible activité de l'année. La troisième, la dernière semaine de novembre, et dans une moindre mesure les premières semaines d'avril correspondent quant à elles aux plus gros pics d'activité. Ces résultats sont cohérents avec ceux de [Rautureau et al., 2011], qui montre que la dernière semaine de novembre 2005 présente la plus forte activité, et la 3ème semaine d'août présente l'activité la plus faible ;

— Lorsque l'on calcule O^V chaque jour, l'activité est plus forte en début de semaine, et décroit au cours de la semaine jusqu'à atteindre son niveau minimal le week-end (ce qu'on l'on observe également sur le réseau d'élevages porcin allemand [Belik et al., 2015]) ;

La figure 2.2 présente à gauche les valeurs prises par O^V et de O^E lorsqu'ils sont calculés chaque mois, des années 2005 à 2015. Autrement dit, on effectue 132 mesures de O^V et de O^E, avec t_i le premier jour et t_f le dernier jour du mois considéré. On observe sur cette figure et sur le tableau 2.1 une tendance globale dans l'évolution de O^V et O^E. On remarque une diminution d'environ 28% du nombre d'exploitations sur les 11 ans étudiés, tout comme [Dutta et al., 2014]. Cette diminution s'accompagne d'une baisse d'environ 1% des animaux échangés, soit environ 21% de liens temporels et 37% de liens statiques. A noter qu'entre 2005 et 2009, période d'étude de [Dutta et al., 2014], la baisse du nombre de liens statiques est d'environ 24%, proche de la valeur de 23% annoncée par les auteurs. Ainsi, la diminution du nombre d'exploitations a un impact important sur le nombre de liens statiques, et dans une moindre mesure le nombre de liens temporels. Au contraire, le nombre d'animaux échangés n'est quasiment pas impacté. Il est donc nécessaire que certains élevages aient augmenté leur nombre d'animaux échangés, ce qui masque la baisse due à la disparition de certains élevages. Cette hypothèse concorde avec l'explication de [Dutta et al., 2014], qui impute la diminution du nombre des exploitations à leur fusion.

2.3.2 Distribution de degré

La définition 1.6 des degrés est à retrouver au chapitre précédant, partie 1.2.3. Ces mesures permettent non seulement d'observer la variété des comportements des exploitations au sein du réseau, mais servent également à caractériser à l'échelle des nœuds leur risque potentiel d'exposition à un pathogène en cas de crise sanitaire.

Nous observons que les allures des distributions du degré total, sortant, et entrant sont stables au cours des années (voir en annexes 7), ce qui est une caractéristique de la BDNI [Dutta et al., 2014]. La figure 2.3 montre l'exemple de l'année 2015, en cumulative inverse. Comme de nombreuses études l'ont mis en évidence en France [Rautureau, 2012] comme à l'étranger (par exemple [Mweu et al., 2013] sur le réseau Danois), les distributions sont hétérogènes. On observe que la plupart des exploitations échangent un faible nombre de bovins, alors qu'une petite fraction du réseau en échange énormément. Ainsi, la distribution du degré total montre qu'un grand nombre d'exploitations (94%) est impliqué dans 1 à 20 échanges en 2015. Un nombre beaucoup plus restreint (1%) est impliqué dans plus de 100 échanges en 2015. Ces nœuds concentrant les interactions sont ce qu'on appelle des hubs, la limite de 100 fixée dans cet exemple étant arbitraire.

Par ailleurs, on remarque que les distributions des degrés entrant et sortant n'ont pas la même allure : les nœuds de faible degré sortant sont en proportion plus élevée que ceux de faible degré entrant. Les valeurs maximales de degré sortant et entrant ont de plus un écart d'un ordre de grandeur : le degré entrant maximal est 5 fois plus élevé que le degré sortant maximal. Les activités d'achat représentent ainsi la majeure partie des échanges entre exploitations.

Lorsque l'on étudie le type d'exploitation des hubs, on remarque que la proportion de marchés et de centres de rassemblement est très élevée, par rapport à leur proportion dans le réseau. Par exemple, ils représentent respectivement 1 et 10% des exploitations ayant un degré supérieur à 100, alors qu'ils ne représentent que $0,04$ et $0,8$% des exploitations dans le réseau de 2015. De plus, les 100 exploitations de plus forts degrés entrant et sortant sont exclusivement des nœuds de type marché ou centre : 10 marchés et 90 centres sont dans les 100 nœuds de plus fort degré entrant, et 15 marchés et 85 centres sont parmi ceux de plus fort degré sortant. En termes de pourcentage, ce sont donc 3% des centres, et 11 à 17% des marchés, qui sont classés parmi les 100 nœuds de plus fort degré.

Dans le chapitre 4, nous verrons les résultats de l'utilisation des degrés comme stratégies de lutte, permettant de cibler les nœuds les plus importants (parce que hubs) dans la dynamique de diffusion.

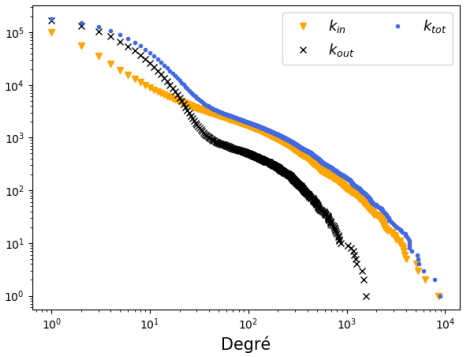


FIGURE 2.3 – Distributions cumulatives inverses des degrés total, entrant et sortant de 2015. Échelles logarithmiques.

2.3.3 Dynamique d'activation des nœuds et des liens

On s'intéresse ici à la question de la répartition des premiers moments d'activation des nœuds et des liens sur une période de temps donnée : ces moments sont-ils tous concentrés dans un intervalle donné ou sont-ils répartis au cours du temps ? Pour faciliter le croisement des résultats avec l'analyse des diffusions effectuée en chapitres 4 et 5, nous divisons au préalable les données en flots de liens de durée annuelle.

Nous étudions donc les moments dans l'année à partir desquels les nœuds et les liens s'activent pour la première fois. Cela revient à observer les ensembles de nœuds et liens distincts sur une période de temps de plus en plus étendue. Dans la figure 2.4, nous calculons O^V, O^V_{in}, O^V_{out}, et O^E (définition 2.1) pour une période de temps $[t_i, t_f]$, où t_i est fixé au $1er$ janvier 2015 et t_f varie entre $t_i + 1$ et le 31 décembre 2015.

On observe des allures différentes selon que l'on mesure O^V_{out} ou O^V_{in} en fonction t_f. D'une part, la valeur de O^V_{out} augmente très rapidement pour $t_t < 50$ jours, puis sa croissance s'infléchit, jusqu'à poursuivre une augmentation faible à partir de 150 jours. Sur l'année 2015, 80% des nœuds ayant une activité sortante vendent pour la première fois très tôt dans l'année (avant 50 jours). À partir de cette date, l'apparition de nœuds s'activant pour la première fois pour vendre demeure faible.

D'autre part, O^V_{in} augmente de manière presque linéaire en fonction de t_f : les nœuds ayant une activité entrante en 2015 s'activent donc tout au long de l'année. On peut donc supposer qu'une maladie pourrait infecter tout au long de l'année des nœuds sains. Ainsi sa dynamique d'infection suivrait l'apparition de nouveaux nœuds ayant une activité entrante. Nous reviendrons sur cette idée dans le chapitre 6.

De même, la figure 2.4 nous montre que O^E croît régulièrement au cours du temps, de

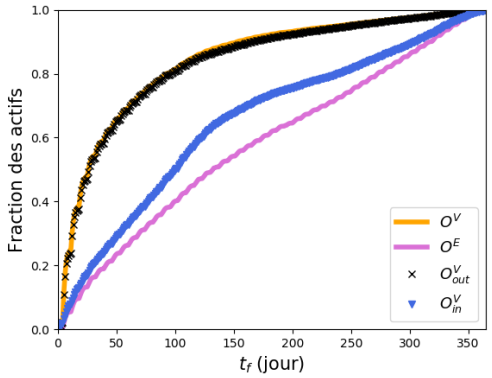


FIGURE 2.4 – Dynamique d'activation des nœuds (ligne orange), des nœuds achetant des animaux (triangles bleus), des nœuds vendeurs (croix noires), et des liens (ligne violette) en 2015. Les résultats sont normalisés respectivement par le nombre de nœuds actifs total, le nombre de nœuds acheteurs actifs, le nombre de nœuds vendeurs actifs, et le nombre de liens actifs en 2015.

manière qualitativement semblable à l'évolution de O^V_{in}. Il existe donc une apparition régulière de nouveaux liens statiques au cours du temps. Ainsi, il semblerait que la progression de la maladie sur le réseau puisse se faire tout au long de l'année, de par le renouvellement des liens pouvant être atteints.

2.3.4 Asymétrie des interactions

Le but de cette mesure est de savoir à quel point un nœud achète ou vend ses animaux aux mêmes exploitations.

[Dutta et al., 2014] observe que le réseau d'élevages français est fortement asymétrique, ce qui veut dire qu'un lien d'une exploitation i vers une exploitation j n'implique pas forcément un lien de j vers i. Un premier aperçu de cette propriété peut être obtenu en comparant le nombre d'exploitations achetant des animaux par rapport au nombre de vendeurs (voir tableau 2.1). Ainsi selon les années, entre 94 et 95% des nœuds vendent des animaux chaque année, contre 58 à 59% qui introduisent un nouvel animal sur leur exploitation.

De plus, la réciprocité (r) peut être utilisée pour mesurer avec plus de précision l'asymétrie du réseau. Elle est définie comme le nombre de liens réciproques (noté $m_{\leftrightarrow}$) sur le

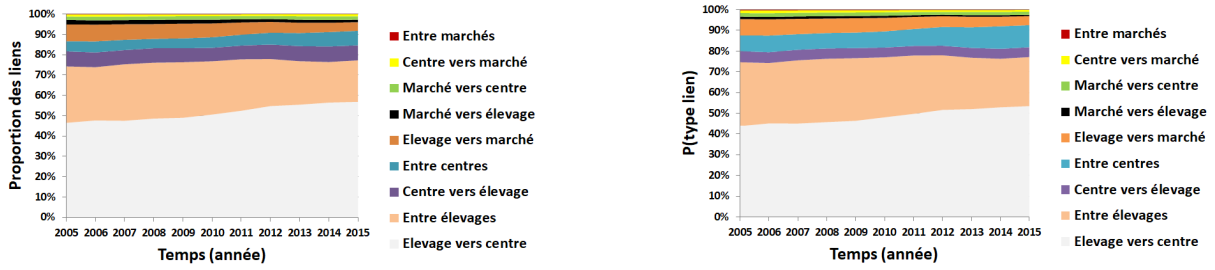
nombre total de liens m, mesurés au cours d'une période donnée :

$$\begin{aligned} &Soit\ G = (V, E) \\ &On\ définit\ m_{\leftrightarrow} = |\{(u, v) \in E, \exists (v, u) \in E\}| \\ &et\ r = \frac{m_{\leftrightarrow}}{m} \end{aligned}$$

Les valeurs de ce taux se situent entre $0,10$ et $0,13$ pour des mesures effectuées chaque année (tableau 2.1), et sont du même ordre de grandeur pour des mesures mensuelles : de $0,06$ à $0,09$. Il existe donc une directionnalité des flux : pour un nœud donné, les bovins achetés ne seront pas vendus aux mêmes exploitations. Les flux entrant et sortant étant donc asymétriques, il est nécessaire de prendre en compte l'orientation des liens lorsque l'on représente les échanges de bovins par un réseau. De plus, cette propriété aura des conséquences directes sur la taille des potentielles épizooties, et sur l'évaluation du niveau de risque auquel serait exposée chaque exploitation. Par exemple, une épidémie a plus de chances d'infecter les nœuds acheteurs.

2.3.5 Fréquence des interactions entre types d'exploitation

On s'intéresse maintenant à la fréquence des liens dans les données, selon leur nature (figure 2.5a et tableau 2.2). La nature des liens est définie en fonction du type des nœuds d'origine et de destination des échanges (élevage, marché, ou centre de rassemblement).



(a) Fréquences mesurées (b) Probabilités

FIGURE 2.5 – A gauche : Fréquence d'apparition des types de liens dans la BDNI, de 2005 à 2015. A droite : Probabilité d'apparition des types de liens dans la BDNI, sur la même période.

Comme dans [Dutta et al., 2014], on observe que les types de liens les plus fréquents sont les liens *élevage vers centre de rassemblement*, suivi des échanges entre élevages.

De 2005 à 2009, la proportion des échanges du type *élevage vers centre de rassemblement* passe de $46,5$ à 49%, puis atteint $56,8\%$ en 2015, soit une augmentation totale de 10 points,

plus particulièrement marquée des années 2009 à 2012. De 2005 à 2009, la proportion des échanges entre élevage reste stable à environ 27%, puis chute pour atteindre près de 20%, soit une diminution d'environ 7 points. Les échanges de type *centre vers élevage* restent quant à eux en proportion environ constante. Toutefois, on observe une augmentation des échanges entre centres, qui passent de 5 à 7% de 2005 à 2015, ce qui rapproche ce type de liens du type *centres vers élevage* en terme de fréquence d'apparition dans les données. Ainsi, on peut penser que les élevages interagissent de plus en plus avec les centres de rassemblement pour échanger leurs animaux, au détriment des interactions directes entre élevages. De plus, on peut supposer que les animaux sont de plus en plus échangés entre centres de rassemblement avant d'être vendus à un élevage, et passent donc plus de temps dans ce type d'exploitation que par le passé. Or, dans le cas d'une maladie à fort potentiel infectieux, les risques de contamination des animaux sont importants dans les centres, ces structures rassemblant un grand nombre d'animaux de diverses origines dans un même lieu. On peut supposer que l'augmentation de la part de ces types d'échanges s'accompagne d'un risque accru de propagation de maladies.

Concernant les autres types d'échanges, leur part reste stable, à l'exception des échanges des élevages vers les marchés. En effet, le pourcentage de ce type de liens restait relativement stable entre 2005 et 2010 (avec $8,3$ et $7,2\%$ des échanges), mais a ensuite réduit de moitié, atteignant $4,3\%$ en 2015. De même, les échanges des marchés vers les élevages voient leur fréquence passer de $2,2$ à 2% de 2005 à 2009, puis atteindre $1,2\%$ en 2015. Il faudrait suivre cette évolution pour voir si les élevages tendent à changer leur comportement d'échange, au profit des centres de rassemblement et au détriment des marchés.

Dans la figure 2.5b, nous représentons la probabilité d'apparition attendue des liens en fonction de leur type, calculée dans les flots de liens de durée annuelle. La probabilité d'apparition $P(x \to y)$ du type de liens temporels *x vers y* est égal à la probabilité que les exploitations de type x aient un lien sortant multiplié par la probabilité que les exploitations de type y aient un lien entrant. Avec V_x et V_y les ensembles de nœuds respectifs des exploitations de type x et y :

$$P(x \to y) = \frac{|\{t,\ tq\ \exists (t,u,v) \in E\ et\ u \in V_x\}|}{|T|} \times \frac{|\{t,\ tq\ \exists (t,u,v) \in E\ et\ v \in V_y\}|}{|T|}$$

La figure 2.5a montre les résultats pour les 11 ans de données. On remarque que les probabilités sont proches des fréquences observées dans les données, avec un écart entre les pourcentages inférieur à 4 points. Seuls les échanges entre centres de rassemblement et des centres vers les élevages peuvent être considérés comme mal approchés par le calcul de leur probabilité :

— la fréquence du type *entre centres* se situe entre $4,6$ et $7,1\%$, alors que leur probabilité théorique est comprise entre $7,4$ et $10,9\%$, soit une probabilité environ 60% plus élevée que la fréquence observée ;

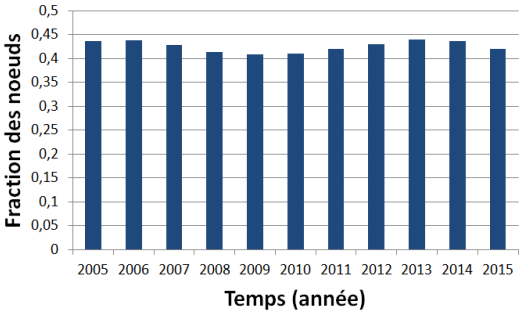


FIGURE 2.6 – Évolution au cours des ans de la taille de la plus grande composante fortement connexe des graphes annuels.

— la fréquence du type *centre vers élevage* se situe entre $6,6$ et $7,7\%$, alors que leur probabilité théorique est comprise entre $4,6$ et $5,3\%$, soit une probabilité plus faible d'environ 30%.

À l'exception des interactions de type *entre centres* et *centre vers élevage*, les valeurs proches de fréquence et de probabilité théorique indiquent que les activités entrantes et sortantes de chaque type de nœud sont suffisantes pour évaluer la fréquence des liens de chaque type.

2.3.6 Étude de la plus grande composante connexe

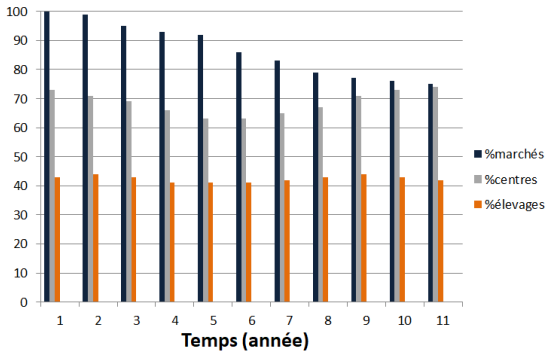
Nous avons vu dans la partie 1.2.3.1 la définition des composantes fortement connexes. Nous avons évoqué également l'existence d'une composante fortement connexe de taille très supérieure à celle des autres composantes du graphe, que l'on appelle la composante fortement connexe géante (GSCC). On observe sur l'exemple de 2015, que la composante fortement connexe géante comprend 74 218 nœuds, la deuxième plus grande en comportant 284. Seul 1% des nœuds font partie d'une composante fortement connexe de plus de deux nœuds.

La figure 2.6 montre l'évolution des tailles des GSCC annuelles au cours des années. On observe que leur taille est relativement stable, avec entre 40 et 45% des nœuds du réseaux étant intégrés aux GSCC.

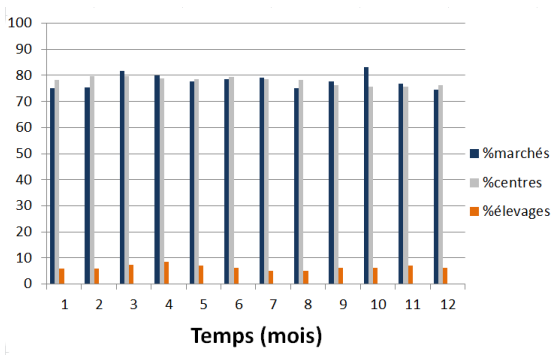
• **Proportions des marchés et des centres de rassemblement**

Nous nous interrogeons plus spécifiquement sur les types d'exploitations représentés dans la GSCC. L'étude de sa composition donne en effet des informations supplémentaires sur les caractéristiques des nœuds jouant un rôle majeur dans la propagation.

Pour mener à bien cette étude, on utilise des séquences de réseaux de différentes durées.



(a) Séquence de graphes annuels 2005 à 2015



(b) Séquence de graphes mensuels de 2015

FIGURE 2.7 – Nombres de nœuds de chaque type dans la GSCC, par rapport à leur nombre étant actif sur les *snapshots* considérés.

La GSCC de chaque réseau de la séquence est calculée, et sa composition est reportée dans le tableau 2.1.

On étudie tout d'abord la séquence de réseaux annuels. On remarque que si les élevages représentent la majeure partie des nœuds, le pourcentage de centres et de marchés dans les GSCC annuelles est légèrement plus élevé que leur pourcentage dans la population : les centres représentent environ 1% des nœuds et les marchés $0,07\%$, contre environ $0,8$ et $0,04\%$.

On observe que la grande majorité des nœuds de ces deux types sont présents dans les GSCC : d'environ 70 à 100% des marchés, et environ 70% des centres en font partie, contre 40% des élevages (figure 2.7a). Si la mesure des GSCC annuelles souligne l'importance du rôle des marchés et des centres de rassemblement dans le réseau d'échanges, elle montre surtout que les GSCC annuelles incluent une grande partie des nœuds du réseau (entre 41 et 44%), indépendamment de leur type.

Ensuite, on observe que le nombre de marchés présents dans les GSCC sur le nombre total de marchés actifs dans l'année diminue au cours du temps, passant de 100% à environ 70% (figure 2.7a). Ceci confirmerait l'hypothèse émise lors de l'étude de la fréquence des liens selon leur type, concernant l'importance décroissante des marchés dans les échanges d'animaux. Le nombre de centres de rassemblement dans les GSCC par rapport à leur nombre actif chaque année oscille quant à lui entre un peu plus de 70% et environ 65% (figure 2.7a), ce qui ne permet pas de conclure quant à un changement de leur importance.

On étudie maintenant la séquence de réseaux mensuels. Tout comme pour les GSCC annuelles, la majorité des nœuds des GSCC mensuelles sont de type élevage. Ils représentent en moyenne 90% des nœuds. Les proportions de centres de rassemblement et de marchés

y sont cependant plus élevées que dans les GSCC annuelles, avec pour les centres entre 7 à 16% des nœuds, et pour les marchés $0,5$ et 1% des nœuds.

Selon les mois, le nombre de marchés faisant partie de la GSCC par rapport à leur nombre dans le réseau mensuel correspondant est compris entre 75 et 83% (figure 2.7b). De même, le pourcentage de centres de rassemblement est compris entre 76 et 80%. Ces pourcentages sont proches de ceux observés dans la séquence de graphes annuels. Au contraire, le nombre d'élevages inclus dans les GSCC mensuelles par rapport à leur nombre total dans le réseau est très différent des valeurs observées pour la séquence de graphes annuels : les pourcentages sont compris entre $4,7$ et $8,8$% sur les *snapshots* mensuels, contre environ 40% sur les *snapshots* annuels. Il est possible que les élevages n'effectuent sur un mois qu'un seul type de transfert, entrant ou sortant, alors que sur une durée d'un an ils effectuent les deux.

À la vue des résultats pour les GSCC annuelles et mensuelles, on peut supposer qu'une durée d'un mois est suffisante pour qu'un nombre élevé de marchés et de centres soient intégrés à la GSCC de par leur activité importante, alors qu'une telle durée est trop courte pour qu'il en soit de même avec les élevages. Ceci est cohérent avec l'idée que les marchés et les centres jouent un rôle très important dans le réseau d'élevage, en permettant de le connecter très rapidement.

2.3.7 Distribution des temps inter-contacts

Dans cette partie, on s'intéresse aux durées qui s'écoulent entre chaque interaction, c'est-à-dire le temps inter-contact (qu'on retrouve en anglais sous l'appellation de *inter-event time*).

On calcule la distribution des τ_i^u (voir définition 1.2) pour tout $u \in V$. L'étude de cette distribution permet d'observer d'éventuelles régularités dans le rythme des échanges, ainsi que de voir si le rythme des échanges est semblable pour tous les nœuds ou leur est totalement propre. Pour une meilleure lisibilité, on étudie les distributions des temps inter-contact des données découpées par année, c'est-à-dire qu'on utilise des flots de liens de durée annuelle. De ce fait, il convient d'enlever les biais générés par la découpe des données en années.

- **Méthode de suppression du biais**

L'extraction de la BDNI demeure un enregistrement borné dans le temps. De ce fait, il existe des effets de bord, dont on veut s'affranchir.

Prenons par exemple un an de données et un nœud interagissant tous les six mois. Si sa dernière interaction se produisait à une date ultérieure à la moitié de l'année, la

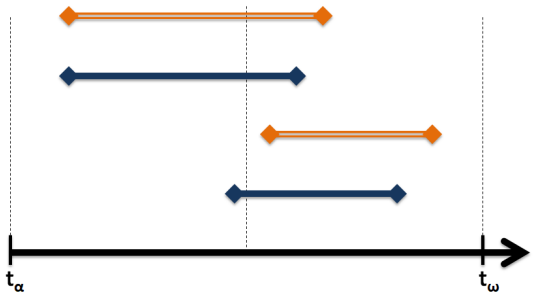


FIGURE 2.8 – Illustration des temps inter-contacts conservés (bleu foncé, trait plein) et ceux supprimés (orange, trait vide) par la *create-based method*.

suivante ne serait pas prise en compte. En effet, la fenêtre temporelle des données n'est pas assez longue pour enregistrer l'interaction suivante et donc comptabiliser le temps inter-contact. Au contraire, son temps inter-contact serait pris en compte dans la distribution si la dernière interaction du nœud datait de la première partie de l'année, et que la suivante se produisait donc dans la seconde moitié. Cet exemple illustre le fait qu'il existe une sous-estimation dans la distribution des durées de temps inter-contact supérieures à la moitié de la durée des données. En conséquence, le nombre des temps inter-contact inférieurs à la moitié de la durée des données est surestimé. Ainsi, l'analyse de la distribution des temps inter-contact est biaisée si aucune méthode de traitement du biais n'est utilisée.

Nous utilisons dans cette étude la *create-based method* pour enlever le biais de la distribution [Roselli et al., 2000]. Cette méthode consiste à comptabiliser uniquement les temps inter-contact qui respectent les règles suivantes (figure 2.8) :

— $\tau_i^u < \frac{t_\omega - t_\alpha}{2}$, avec t_α le temps initial et t_ω le temps final du flot de liens ;

— $t_i < t_\omega - \frac{t_\omega - t_\alpha}{2}$; $t_{i+1} < t_\omega$.

Autrement dit, on conserve uniquement les temps inter-contact dont la durée est inférieure à la moitié de la période d'analyse, et débutant dans la première moitié de la période.

- **Analyse des distributions**

Sur un flot de liens donné de durée annuelle, on calcule pour tout u dans V tous les τ_i^u, puis on rassemble tous les résultats des nœuds pour en faire la distribution (figure 2.9 pour l'année 2015 ; voir en annexes pour les résultats des autres années, qui sont similaires). On observe une périodicité dans les valeurs, qui correspond aux semaines : il est plus probable qu'un nœud interagisse avec une périodicité de 7, 14, 21 jours, etc., plutôt qu'avec une périodicité quelconque. Cette périodicité s'accompagne d'une décroissance de la probabilité : les temps inter-contacts longs sont moins probables que les courts. Par exemple, il est plus probable qu'un nœud interagisse tous les 7 jours que tous les 14. Nous observerons au chapitre 5 que cette caractéristique a effectivement un impact sur la diffusion, se traduisant par une accélération des transmissions en début de semaine.

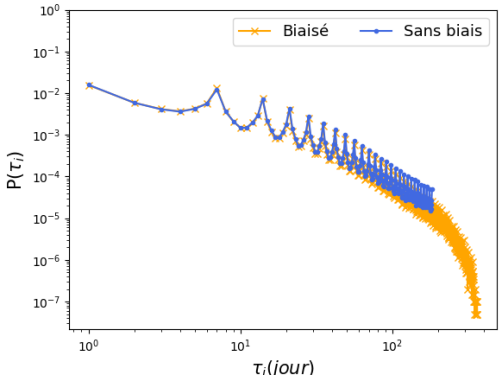


FIGURE 2.9 – Distribution des temps inter-contact entre deux interactions de l'année 2015, avec ou sans correction du biais. Échelles logarithmiques.

- **Temps inter-contact entre un achat et une vente**

Jusqu'à présent, nous mesurions le temps inter-contact entre deux interactions, sans consi-

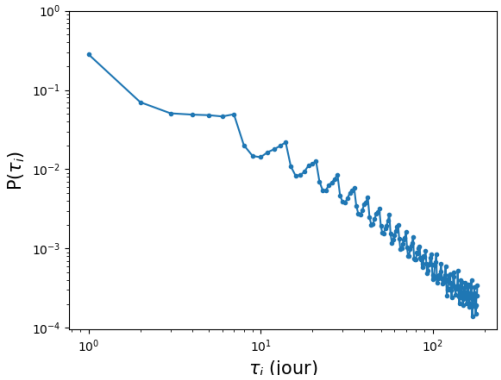


FIGURE 2.10 – Distribution des temps inter-contact entre un achat et une vente pour chaque nœud (année 2015). Ces deux échanges ne doivent pas avoir lieu le même jour. Échelles logarithmiques.

dération pour leur nature : les interactions pouvaient correspondre à un achat comme à une vente. La figure 2.10 montre la distribution des temps séparant un achat de la vente suivante, pour chaque nœud du réseau. Les biais sont toujours enlevés grâce à la *create-based method*. Nous en tirons les mêmes conclusions, à savoir que les distributions présentent une décroissance avec une périodicité correspondant aux semaines. Ainsi, les exploitations mènent leurs activités d'achat et de vente selon un schéma similaire (cf chapitre 5).

Chapitre 3

Diversité des tailles de propagations

Sommaire

Lorsque l'on étudie la vulnérabilité d'un réseau à la propagation de maladies, on cherche à évaluer quelle fraction des nœuds risque d'être atteinte. Cependant, cet impact est grandement dépendant du lieu de départ de la diffusion. En effet, il existe une grande diversité d'activité, de degré des nœuds, et tous ne se trouvent pas sur des chemins ayant la même importance dans le réseau. De par leurs interactions, ces nœuds s'organisent de plus en différentes structures, qui peuvent interagir préférentiellement entre elles. Les propriétés des nœuds de départ ont donc une grande importance sur la fraction finale des nœuds qui pourra être atteinte.

Connaître la structure du réseau permet donc d'obtenir une première indication sur l'impact qu'aurait une propagation, en fonction de son point de départ. Pour étudier cet impact, différentes mesures ont été proposées dans la littérature, que nous présenterons puis appliquerons aux données de la BDNI. Ensuite, nous détaillerons un protocole pour mesurer les tailles des propagations sur un flot de liens, dans des conditions les plus proches possibles des mesures statiques détaillées plus tôt. Le but est ainsi d'étudier dans le contexte de la BDNI, l'impact de la temporalité sur ce type de mesure. Enfin, nous étudierons les propagations au moyen de cette mesure temporelle, et nous conclurons quant à sa valeur-ajoutée par rapport aux mesures statiques.

3.1 Analyse dans le cas statique

3.1.1 Représentation en nœud papillon d'un graphe orienté

Dans le premier chapitre (partie 1.2.3.1), nous avons défini la notion de composante connexe dans le cas d'un graphe orienté ou non orienté : une composante (fortement) connexe est un groupe maximal de nœuds tel qu'il existe un chemin (orienté) entre toutes les paires de nœuds le composant. L'existence d'une composante géante, c'est-à-dire de taille nettement supérieure aux autres composantes connexes, est une caractéristique fréquente des réseaux réels (cf partie 2.3.6). Une maladie débutant en son sein aura donc un impact plus grand qu'une autre débutant dans une petite composante. L'étude de la structure du graphe, et en particulier de cette composante géante, est donc d'un grand intérêt pour évaluer l'impact potentiel qu'aurait une propagation en fonction de son lieu de départ dans le réseau.

On note GWCC (de l'anglais *giant weakly connected component*) et GSCC (*giant strongly connected component*) les composantes connexe et fortement connexe de plus grandes tailles.

La représentation des graphes orientés en nœud papillon (de l'anglais *bow-tie structure*) a été mise au point pour décrire leur structure [Broder et al., 2000]. Si le sujet d'étude des auteurs est la description du web, l'organisation en nœud papillon se retrouve dans tous les graphes orientés, quelle que soit leur nature. Un tel réseau s'organise en effet autour de trois grandes structures, voir figure 3.1 :

- un cœur central (*central core*), correspondant à la GSCC ;
- une composante entrante (*in-component*), qui correspond à l'ensemble des nœuds pouvant atteindre la GSCC, mais qui ne peuvent pas être atteints par ces mêmes nœuds :
  $IN = \{u \in V \setminus GSCC,\ \exists v \in GSCC\ tq\ u \rightsquigarrow v\}$;
- une composante sortante (*out-component*), regroupant tous les nœuds pouvant être atteints par ceux de la GSCC, mais qui ne peuvent pas les atteindre :
  $OUT = \{u \in V \setminus GSCC, \exists v \in GSCC\ tq\ v \rightsquigarrow u\}$;

De plus, le réseau peut comporter :

- des tubes, reliant les composantes entrante et sortante sans passer par le cœur :
  $tube = \{w \in V \setminus GSCC,\ \exists u \in IN,\ \exists\ v \in OUT\ tq\ u \rightsquigarrow w \rightsquigarrow v\}$;
- des vrilles (*tendrils*), éléments soit atteints par la composante entrante, soit atteignant la composante sortante :
  $tendril = \{u \in V \setminus (GSCC \cup IN \cup OUT \cup tube),\ \exists v \in (IN \cup OUT \cup tube)\ tq\ u \rightsquigarrow v\ ou\ v \rightsquigarrow u\ \}$.
- des composantes connexes isolées, c'est-à-dire sans chemin reliant un de leurs nœuds

à une autre des composantes ci-dessus.

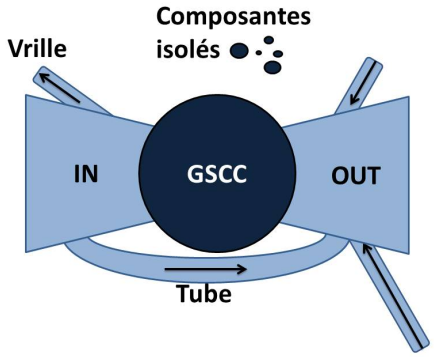


FIGURE 3.1 – Représentation en nœud papillon d'un graphe orienté, exhibant une composante géante.

3.1.2 Affiner la mesure avec l'utilisation d'un modèle susceptible-infecté

Dans la littérature, on retrouve les mesures des tailles des GWCC (voir par exemple [Bajardi et al., 2011, Büttner et al., 2013]) et des GSCC (voir par exemple [Rautureau et al., 2011, Dutta et al., 2014]) pour évaluer la taille maximale que peut atteindre une propagation sur un réseau. En effet, selon [Robinson et al., 2007] :

— la taille de la GWCC représente la borne maximale que peut prendre le pire cas de propagation. Autrement dit, la plus grande cascade atteindra un nombre de nœuds inférieur à la taille de la GWCC ;

— la taille de la GSCC représente la borne minimale du pire cas de propagation. En d'autres termes, la plus grande cascade mesurée atteindra au moins un nombre de nœuds égal à la taille de la GSCC.

Cependant, ces grandeurs délimitent une fourchette large de taille de propagation maximale : par exemple pour 2015, la GSCC rassemble 42% des nœuds, contre 98% pour la GWCC, soit plus du double. Pour obtenir plus de précision quant à la mesure des tailles de propagations, dans le cas d'un graphe orienté, il faut alors se tourner vers l'utilisation d'un modèle SI avec un taux d'infection de 1, comme dans [Belik et al., 2015] ou [Lentz et al., 2016].

En effet, le nombre de nœuds atteints par ce modèle de propagation constitue une mesure déterministe, permettant d'évaluer la connectivité du réseau : pour un nœud donné, choisi comme source de la propagation, tous les nœuds vers lesquels il existe un chemin (défini en 1.3) finissent par être atteints. Le nombre de nœuds atteints est donné par la

taille de la cascade au départ du nœud choisi (voir partie 1.2.3.1). Pour rappel, dans un graphe orienté $G = (V, E)$, on note $\mathscr{C}(v)$ la cascade au départ de $v \in V$, et $V_{\mathscr{C}(v)}$ l'ensemble des nœuds de la cascade. $|V_{\mathscr{C}(v)}|$ donne donc la taille de la cascade $\mathscr{C}(v)$. Avec la mesure des tailles des cascades, on obtient une évaluation précise de l'impact du nœud de départ sur les tailles des propagations. De plus, elle permet d'évaluer le pire cas de propagation, c'est-à-dire la propagation de taille maximale : $max(|V_{\mathscr{C}(v)}|)_{v \in V}$.

3.1.3 Application au cas de la BDNI

La BDNI, une fois modélisée par un graphe orienté (et statique) dont les nœuds sont les exploitations et les liens les échanges d'animaux (cf chapitre 2), se prête naturellement à une analyse de la structure en nœud papillon. D'abord, on calcule la taille de la GSCC et de la GWCC. Puis, on mesure le nombre de nœuds atteints par un modèle SI sur le graphe $G = (V, E)$ au départ de chaque nœud, c'est-à-dire les $|V_{\mathscr{C}(v)}|$ pour tout v dans V.

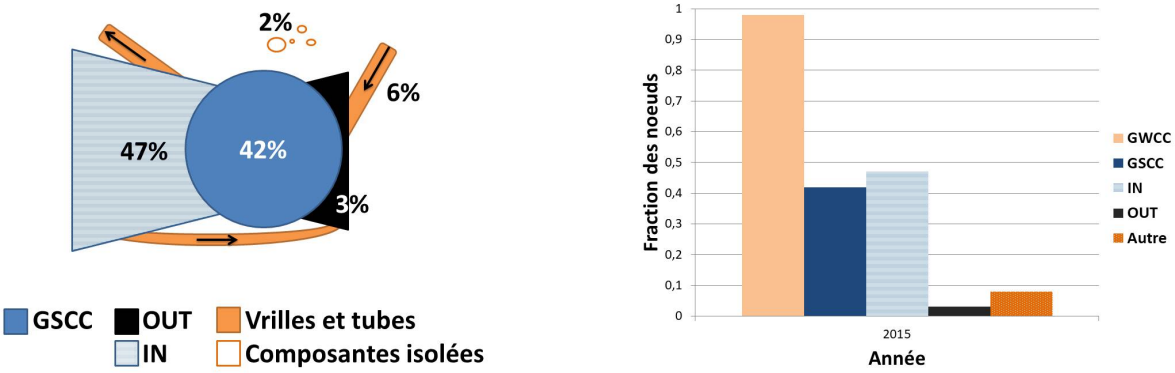


FIGURE 3.2 – Fraction des nœuds du graphe de 2015, appartenant aux différentes structures du nœud papillon.

La figure 3.2 représente l'organisation du graphe de l'année 2015 de la BDNI :

— une maladie débutant au sein du cœur central se propagera à tous les autres nœuds de la GSCC (74 217, soit 42% des nœuds actifs en 2015) plus ceux de la composante sortante (5258 soit 3% en 2015) ;
— environ 47% des nœuds appartiennent à la composante entrante. Avec les nœuds de la GSCC, ils présentent le plus grand risque de propagation pour le réseau ;
— il reste donc 6% des nœuds de 2015, qui appartiennent aux tubes ou aux vrilles. Ils sont sources de petites cascades ;
— près de 98% (172 706) des nœuds actifs en 2015 appartiennent à la GWCC ;
— les 2% des nœuds restants sont isolés dans de petites composantes connexes. Une propagation débutant depuis l'une d'entre elles aura donc un impact très limité sur

le réseau.
La description des autres années est à retrouver en annexes. On constate que la répartition des nœuds entre les différentes composantes du nœud papillon est très stable au cours du temps.

Par ailleurs, dans [Lentz et al., 2016], les auteurs utilisent une représentation en nœud papillon pour décrire le réseau d'échanges de porcins en Allemagne. Ils trouvent une répartition qualitativement similaire des nœuds entre les différentes structures. Ainsi, la structure statique des réseaux d'échanges d'animaux se ressemblerait, même pour des pays et des espèces différents.

L'utilisation de la représentation en nœud papillon permet de caractériser la capacité des nœuds à être source d'un propagation de taille importante, en fonction de leur localisation.
Dans une représentation en graphe, elle permet d'observer quelles proportions de la population risquent d'être à l'origine de grandes propagations (89% des nœuds en 2015, pour la BDNI), ou de propagations de taille plus modeste.

3.2 Analyses dynamiques

3.2.1 État de l'art

Les méthodes d'estimation des tailles de propagation sur un graphe (GWCC, GSCC comme modèle SI) ne prennent pas en compte l'ordre chronologique des interactions dans leur calcul. La BDNI donnant accès à ce type d'information, on peut s'interroger sur la nécessité de se tourner vers des mesures prenant en compte la temporalité pour le calcul du nombre de nœuds atteints. En effet, lorsque l'information temporelle est négligée, une propagation de A vers B puis de B vers C est considérée comme équivalente à une propagation de B vers C puis de A vers B, ce qui n'est pas le cas par exemple si A était infecté : dans le premier cas, les trois exploitations finissent par contracter la maladie, alors que dans le deuxième cas, seule B est contaminée par A. De plus en plus d'études soulignent l'impact de la prise en compte de l'information temporelle sur la précision de ces mesures. Par exemple, [Vernon and Keeling, 2009] montrent que la fréquence et l'ordre des interactions s'avèrent importants lors de la simulation de phénomènes infectieux, de par la dynamique intrinsèque aux données de mouvements d'élevage. Dans [Dubé et al., 2008], les auteurs montrent quant à eux que mesurer les tailles de SCC conduit à surestimer la propagation d'une maladie, en raison de la non prise en compte de l'ordre temporel des interactions. Les auteurs conseillent donc d'utiliser des mesures telles que la chaîne d'infection sortante (cf partie 1.2.3), permettant de tirer parti de l'information temporelle

contenue dans les données. Cette partie s'inscrit dans cette optique : elle compare, dans le cas de la BDNI, des mesures statique et dynamique pour évaluer les tailles de propagation.

3.2.2 Notre protocole

Pour mesurer les tailles de propagation en contexte temporel, nous choisissons de mesurer les chaînes d'infection sortantes. Cette mesure est équivalente au nombre de nœuds atteints par un modèle SI avec un taux d'infection égal à 1 respectant l'ordre temporel des liens, autrement dit, un SI sur un flot de liens. Ainsi, on peut la voir comme une généralisation de la méthode décrite dans la partie 3.1.2, qui compte les nœuds atteints dans un graphe par un modèle SI. Nous présentons maintenant un protocole pour comparer les résultats de ces deux mesures, dans des conditions les plus proches possibles.

Lors de la transformation des données en séquences de graphes, chaque graphe représente l'agrégation d'interactions ayant lieu au cours de périodes de même durée. Ainsi, même si les graphes ne conservent pas l'information temporelle, la taille des propagations est dépendante de la durée de la période considérée : la diffusion se fait uniquement via les interactions actives durant cette période. Une plus longue période augmente le nombre d'interactions rencontrées lors de la diffusion et donc la taille de propagation. Par exemple, le plus grand nombre de nœuds atteints dans le graphe annuel de 2015 est de 79 510, alors qu'il est de 13 241 sur les graphes mensuels de 2015. Afin de comparer les tailles de propagation dans des situations les plus proches possible, il convient de fixer une durée de propagation limite pour le calcul des chaînes d'infection. On fixe cette durée pour qu'elle corresponde à celle utilisée pour construire la séquence de graphes étudiée. Autrement, des quantités différentes d'informations seraient considérées ce qui biaiserait la comparaison. C'est pourquoi, nous bornerons les cascades dans le temps, comme [Schärrer et al., 2015] pour le calcul des chaînes d'infection entrantes [1] (voir en partie 1.2.3.1 et la définition 1.5).

Pour résumer, on procède comme suit :
- on part d'un flot de lien : $L = (T, V, E)$, avec $T = [t_\alpha, t_\omega]$;
- on fixe une durée $d = \frac{t_\omega - t_\alpha}{k}$, qui va permettre de découper T en k fenêtres temporelles, avec k un entier supérieur à 1 ;
- on peut donc considérer k flots de liens $L_i = (T_i, V_i, E_i)$, avec $T_i = [t_\alpha + d \cdot i, t_\alpha + d \cdot (i+1)[$, et $E_i = E \cap T \times V \times V$, pour $i \in \{0, 1, \ldots, k-1\}$;
- on construit la séquence de *snapshots* correspondante, $S = (G_i)_{i=0,1,\ldots,k-1}$. Chaque graphe G_i correspond à l'agrégation des interactions du flot de liens L_i ;
- pour chaque i, pour chaque $u \in V_i$, on calcule le nombre de nœuds atteints, à partir

1. Chaîne d'infection entrante = nombre de nœuds pouvant atteindre le nœud étudié, en respectant l'ordre chronologique des interactions. On peut donc considérer que c'est la taille d'une cascade lorsque l'on remonte dans le temps.

de u dans G_i ;

— pour chaque $u \in V$, on sélectionne un temps de départ pour la cascade bornée dans le temps dans le flot L. On veut éviter de choisir un temps de départ dans une période d'inactivité de u (section 2.3.7), ce qui utiliserait une partie de la durée d allouée à la propagation à attendre que le nœud s'active. Pour ce faire, nous tirons aléatoirement un temps d'interaction où le nœud a une activité sortante, c'est-à-dire un temps t tel que $\exists(t, u, v) \in E$ et $t \leq t_\omega - d$. On note $\mathcal{W}$ l'ensemble des couples (t, u) ainsi obtenus.

— on calcule les tailles des cascades (*i.e.* leurs chaînes d'infection sortantes) bornées dans le temps, pour départ (t, u) dans $\mathcal{W}$ (définition 1.5).

Comme présenté dans le chapitre 2, certains nœuds du réseau n'ont aucun mouvement sortant sur une période de temps donnée. De ce fait, ils n'obtiendront pas de chaîne d'infection sortante. La figure 3.3 donne un exemple de diffusion dans un flot de liens et la séquence de graphes correspondante.

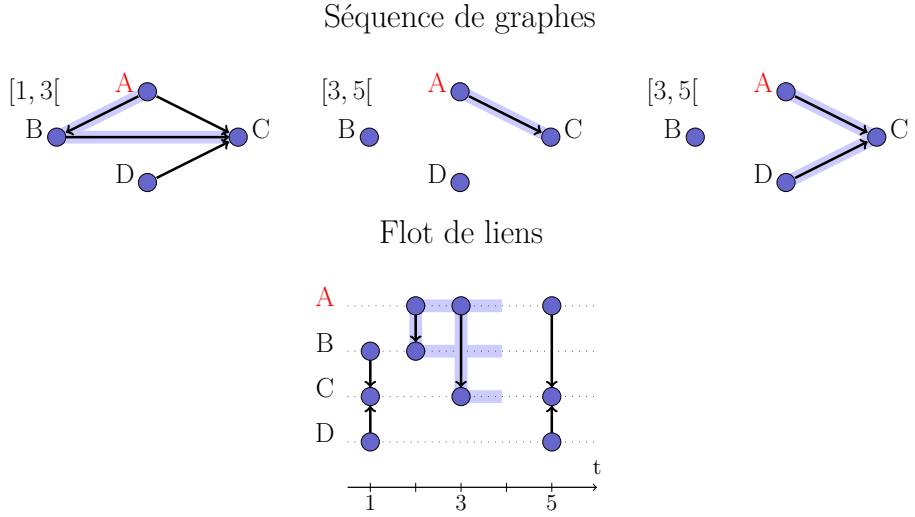
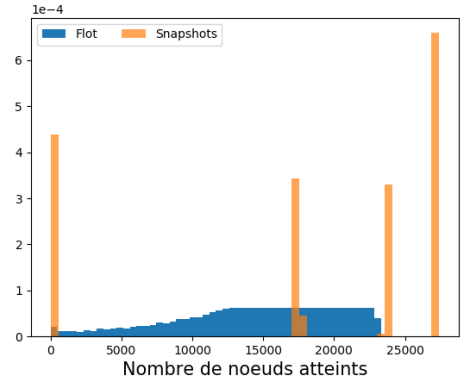


FIGURE 3.3 – Protocole de comparaison de propagations sur une séquence de graphes (en haut), où chaque *snapshot* représente une durée de deux unités de temps, et sur un flot de liens (en bas). Le nœud A est choisi dans cet exemple comme nœud source pour les diffusions. Dans la séquence de graphes, il atteint 2 nœuds dans le premier *snapshot*, et 1 dans les deuxième et troisième *snapshots*. Supposons que le temps d'interaction 2 ait été choisi comme temps de départ de la propagation depuis A dans le flot de liens. La propagation atteint alors deux nœuds, aux temps 2 et 3. Puis, la diffusion s'arrête lorsqu'elle atteint la fin de la période de propagation de deux unités de temps, fixée lors de la construction de la séquence de graphes.

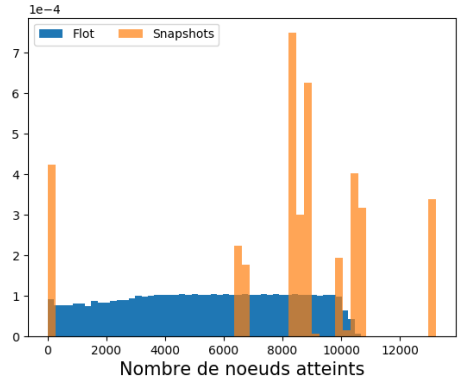
3.2.3 Résultats sur la BDNI

Dans cette partie, nous présentons les résultats de comparaison des cas statique et temporel pour l'année 2015. Nous choisissons des *snapshots* de durée d'un mois et d'un trimestre, comme dans [Dutta et al., 2014]. Par conséquent, nos cascades auront également une durée limitée à un mois, ou un trimestre. Les résultats des autres années sont présentés en annexes et sont qualitativement similaires.

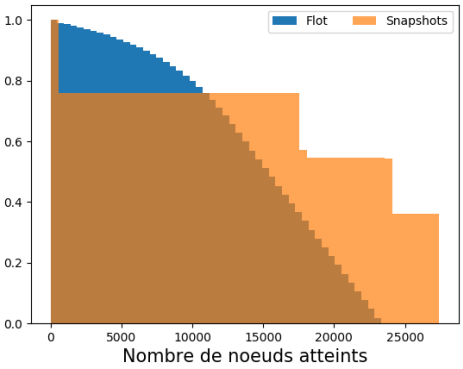
Les figures 3.4a et 3.4b présentent les distributions des tailles de propagation respectivement pour les diffusions d'un mois et trimestrielles. Tout d'abord, on s'intéresse à l'allure des distributions, puis aux tailles de propagation estimées.



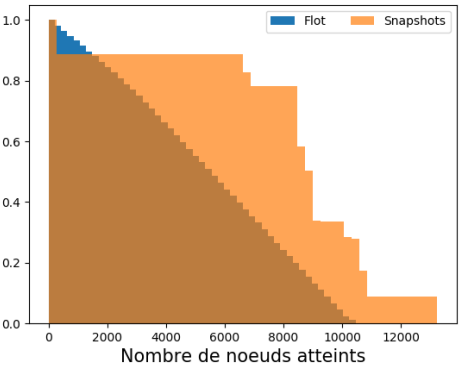
(a) Propagations d'un trimestre



(b) Propagations d'un mois



(c) Propagations d'un trimestre (cumulative inverse)



(d) Propagations d'un mois (cumulative inverse)

FIGURE 3.4 – Distribution du nombre de nœuds atteints par un SI de taux d'infection de 1, sur les graphes trimestriels et mensuels de 2015, en comparaison des propagations d'un trimestre ou d'un mois sur le flot de liens de 2015. Valeurs discrétisées, en 50 intervalles.

On observe une différence de nature entre les distributions : alors que les distributions des chaînes d'infection présentent un continuum entre les valeurs prises, les distributions des tailles de cascades sur les graphes orientés montrent une faible diversité de valeurs. En effet, les distributions cumulatives inverses montrent que les nombres de nœuds atteints sont :

— soit proches de la taille maximale de propagation. Plus de 70% des propagations sur graphes trimestriels (figure 3.4c), et de plus de 80% de celles sur graphes mensuels (figure 3.4d) ont des tailles supérieures à la moitié de la propagation de taille maximale ;
— soit faibles devant la taille maximale (quelques centaines de nœuds atteints, que ce soit pour les séquences de graphes mensuels ou trimestriels).

L'apparition de ces distributions montrant une faible diversité de tailles de cascades est due à l'existence d'une GWCC. On observe que lorsque la durée des *snapshots* augmente, l'impact des propagations débutant au sein des composantes entrantes diminue, par rapport à celui des propagations débutant dans les GSCC. En effet :

— pour une durée d'un mois en 2015, la plus grande propagation, débutant dans la composante entrante, atteint 50% de nœuds en plus qu'une propagation débutant dans la GSCC ;
— pour une durée d'un trimestre en 2015, 36% de nœuds supplémentaires sont atteints par la plus grande propagation, par rapport à celles débutant dans la GSCC ;

En 2015, plus la durée du *snapshot* est faible, plus la variabilité des tailles de propagations débutant dans les composantes entrantes est élevée. Dans le cas des *snapshots* de courte durée, la taille de la GSCC tend à s'éloigner de la taille maximale de propagation.

Au contraire des distributions obtenues sur les graphes, les distributions des chaînes d'infection sortantes sur les flots de liens présentent une grande diversité de valeurs. Autrement dit, la mesure temporelle de la taille des infections permet de détecter toutes les tailles intermédiaires, contrairement à la mesure équivalente sur les graphes. En effet, concernant les propagations d'une durée d'un mois, les cascades de tailles intermédiaires, non détectées sur les graphes, représentent 18% des propagations sur les flots de liens (tableau 3.1a). La proportion des propagations de tailles intermédiaires est même de 31% dans les flots de liens, dans le cas des diffusions d'une durée d'un trimestre (tableau 3.1b). Il existe donc une diversité de tailles de propagation qui n'est pas détectée dans les séquences de graphes. Cette diversité avait également été montrée sur le réseau d'échange de porcins en Allemagne [Lentz et al., 2016]. Cette caractéristique pourrait donc être généralisable à d'autres jeux de données d'échanges d'animaux.

En plus de présenter des distributions d'allures différentes, les figures 3.4a et 3.4b, et le tableau 3.2 nous permettent de comparer l'évaluation des tailles maximales de propagation, lorsque la temporalité est prise en compte ou non. Indépendamment de la durée de propagation étudiée, les SI prenant en compte l'ordre chronologique des interactions

TABLE 3.1 – Comparaison des pourcentages de propagations selon leurs tailles.

(a) Propagations d'un mois

Nombre de nœuds atteints	Moins de 100	Entre 100 et 5000	Plus de 5000
Pourcentage (nombre moyen) des propagations sur un graphe mensuel	12% (11207)	-	88% (82599)
Pourcentage (nombre) des propagations de durée d'un mois sur un flot de liens	8% (10643)	18% (24267)	74% (100756)

(b) Propagations d'un trimestre

Nombre de nœuds atteints	Moins de 200	Entre 200 et 15000	Plus de 15000
Pourcentage (nombre moyen) des propagations sur un graphe trimestriel	9% (12233)	-	91% (125781)
Pourcentage (nombre) des propagations de durée de 3 mois sur un flot de liens	4,5% (7208)	31% (50668)	64% (103562)

mènent nécessairement à des nombres maximaux de nœuds atteints plus faibles que leur équivalent statique.

Ainsi, non seulement négliger l'information temporelle ne permet pas de mesurer précisément les tailles de diffusion et surestime les propagations de tailles intermédiaires, mais en plus, cela entraîne une surestimation de la taille maximale possible de propagation. Comme sur de nombreux réseaux réels, il est essentiel de prendre en compte l'information temporelle pour étudier l'impact potentiel d'une propagation sur la BDNI.

TABLE 3.2 – Comparaison du nombre maximal de nœuds atteints pour des propagations mensuelles et trimestrielles, en nombre absolu, en pourcentage du nombre de nœuds n, et en pourcentage du nombre de nœuds actifs en 2015 ayant un degré entrant supérieur à 1.

	Graphes mensuels	Flot 1 mois	Graphes trimestriels	Flot 3 mois
Taille maximale	13241	10673	27429	23327
en %n	7,5%	6%	15,5%	13%
en %n, $deg_{in}>0$	13%	10,5%	27%	23%

La mesure des tailles des cascades sur les réseaux temporels montre que celles-ci atteignent des nombres très variés de nœuds. Cette diversité ne peut pas être détectée au moyen de mesures statiques, sur les graphes. Les propagations de tailles intermédiaires ne sont notamment pas détectées dans le cas statique.

Conclusion

Utiliser la taille des GSCC comme indicateur des tailles de propagation ne permet pas d'évaluer correctement le pire cas de propagation possible. Si mesurer le nombre de nœuds atteints par un modèle SI répond à cette problématique, les tailles de propagations sont toutefois surestimées. C'est plus particulièrement le cas des propagations de tailles intermédiaires.

Au contraire, mesurer les chaînes d'infection sortantes permet une évaluation plus précise des tailles des propagations de la BDNI, en prenant en compte la temporalité des échanges. Elle permet notamment de mettre en évidence la diversité des tailles de propagations. Notamment, elle montre qu'il existe une grande variété de tailles intermédiaires, entre les valeurs minimales et maximales de chaînes d'infection sortantes.

Chapitre 4

Identification d'éléments clés pour la propagation

Sommaire

Nous avons vu dans le chapitre 2 que tous les nœuds et liens n'ont pas les mêmes caractéristiques. Par exemple, certains nœuds sont impliqués dans de nombreux échanges, et seront donc certainement plus facilement atteints par une propagation que les nœuds échangeant peu. D'autres interagissent par pics d'activité (*bursts*), caractérisés par des temps inter-contacts très courts, ce qui permet à une diffusion d'atteindre rapidement de nouveaux nœuds. Ou encore, certains nœuds sont essentiels dans la structure du réseau, en permettant à des parties autrement isolées d'interagir. Avec ces exemples, nous voyons qu'il existe de nombreuses façons d'être important dans le réseau. Chaque notion d'importance repose sur différentes propriétés des nœuds et des liens. Aussi ne nécessitent elles pas la même information sur les interactions. Par exemple, certaines notions nécessitent d'avoir accès à l'information temporelle sur les échanges. D'autres ne nécessitent qu'une connaissance locale et non temporelle des interactions des nœuds. En fonction des données accessibles et du but recherché, des nœuds et des liens potentiellement différents sont identifiés comme importants par les différentes notions.

Les nœuds et liens identifiés par les notions d'importance sont a priori à cibler spécifiquement par les mesures de surveillance ou de lutte. On peut ensuite évaluer l'efficacité de

ces mesures, par des expérimentations similaires à celles présentées dans le chapitre précédant : le traitement des nœuds et liens identifiés comme importants a-t-il effectivement permis de réduire la taille des propagations ? De plus, quel a été le nombre de nœuds et de liens identifiés et traités pour permettre cette réduction de l'impact des propagations ? En effet, les ressources financières tant qu'humaines sont limitées pour mettre en place des stratégies de surveillance et de lutte [Salathé and Jones, 2010, Dutta et al., 2014]. On ne peut donc pas simuler la lutte en autorisant le traitement des nœuds et liens sans limite de nombre.

De nombreuses études portent donc sur l'amélioration des méthodes d'identification des nœuds et des liens importants pour les processus de diffusion. Nous présentons les réponses apportées par la littérature à ces différents points dans la première partie de ce chapitre. Par rapport aux stratégies existantes, notre but est d'étudier l'impact de la prise en compte de la temporalité des interactions sur l'identification des nœuds et liens importants. Autrement dit, nous souhaitons comparer l'efficacité des méthodes d'identification utilisant l'information temporelle des données par rapport à celles négligeant cette information. La présentation de stratégies temporelles et d'une méthode de comparaison avec les stratégies statiques sont l'objet de la deuxième partie de ce chapitre. Celui-ci s'achèvera par la comparaison de l'efficacité des stratégies temporelles et statiques, et par la conclusion quant à l'importance de prendre en compte l'information temporelle sur les échanges d'animaux.

4.1 État de l'art

Les mesures de lutte contre la diffusion de maladies font référence usuellement à la vaccination, l'abattage, et la mise en quarantaine des animaux d'une exploitation. La place de l'amélioration des stratégies de vaccination est importante dans la littérature, étant donné qu'un grand nombre d'études ont pour sujet la diffusion de maladies humaines. Du point de vue d'un réseau, la vaccination est néanmoins modélisée de la même façon que l'abattage ou la mise en quarantaine : les nœuds identifiés comme importants sont retirés du réseau, et avec eux toutes leurs interactions. En effet, toutes ces stratégies consistent à empêcher le nœud ciblé de transmettre la maladie :

- Dans le cas de l'abattage, le lien avec la suppression du nœud est direct.
- Dans le cas de la quarantaine, le nœud ne peut plus interagir avec le reste du réseau. Il ne peut ni être infecté, ni transmettre la maladie. Modéliser cet isolement est équivalent à supprimer le nœud.
- Dans le cas de la vaccination, il est considéré que celle-ci apporte une immunité permanente aux nœuds traités. Lorsqu'une maladie atteint ces nœuds, ils ne sont donc jamais infectés, et ne peuvent donc jamais transmettre la maladie à leurs voisins. Par rapport à la diffusion de la maladie, les nœuds vaccinés sont des impasses

sur les chemins de propagation. Les supprimer du réseau permet donc de modéliser leur état de nœud vacciné.

Si dans le contexte épidémiologique, on trouve principalement des stratégies de suppression des nœuds, il est également possible de mettre en place des stratégies de suppression des liens, comme nous le verrons par la suite.

Pour supprimer les nœuds ou les liens d'un réseau, une méthode simple serait de les cibler de manière aléatoire. [Magnien et al., 2011] étudient par exemple l'impact de différents scénarios de suppression aléatoire (de nœuds ou de liens) sur la taille de la plus grande composante connexe de différents réseaux réels. Cependant, les réseaux réels sont caractérisés par une hétérogénéité marquée de leurs propriétés : comme nous l'avons vu au chapitre 2, les nœuds présentent une distribution des degrés, des temps inter-contact etc. hétérogènes. De nombreuses études ont souligné l'impact important de ces hétérogénéités sur la mise en place de mesures de lutte [Pastor-Satorras and Vespignani, 2002, Barrat et al., 2008] : elles constatent que supprimer aléatoirement les nœuds d'un réseau réel nécessite de traiter presque tous les nœuds. Par exemple, il est nécessaire de vacciner 95% de la population pour empêcher la diffusion de la rougeole [Anderson and May, 1992], lorsque la vaccination n'est pas ciblée. Le coût de la lutte étant important, des stratégies ciblées et adaptées aux caractéristiques des réseaux réels ont été développées, comme nous allons le voir à présent.

4.1.1 Identification des éléments à cibler

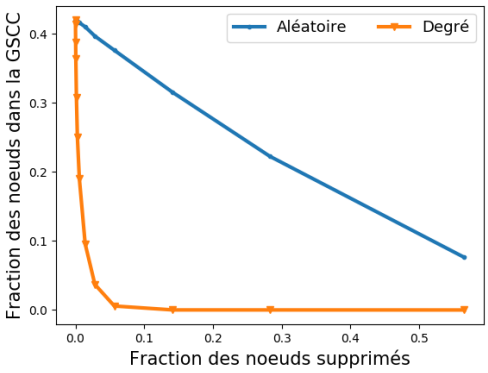


FIGURE 4.1 – Tailles des plus grandes composantes fortement connexes en fonction de la fraction de nœuds supprimés, pour l'année 2015 de la BDNI.

Au lieu de tirer aléatoirement les nœuds à supprimer, le but de ces stratégies est de cibler les nœuds ayant les caractéristiques les plus favorables à la diffusion de maladies. Par exemple, les nœuds ayant le plus d'interactions ont de ce fait plus d'occasions de diffuser

une maladie. La suppression des nœuds de plus fort degré (définition 1.6) est donc une stratégie de ciblage typique [Pastor-Satorras and Vespignani, 2002, Chen et al., 2008]. Sur l'exemple de la figure 4.1, il faut supprimer 60% des nœuds pour que la taille de la plus grande composante fortement connexe soit réduite à 10% des nœuds du réseau, lorsque les suppressions sont menées de manière aléatoire. Au contraire, 3% de nœuds supprimés suffit lorsqu'ils sont choisis selon leur degré décroissant (du plus fort degré au plus faible score). Dans le cas de la suppression de liens, [Isella et al., 2011] proposent de les supprimer selon leur poids (définition 1.7), c'est-à-dire selon le volume des interactions passant par les liens statiques. Dans un cadre temporel, une façon de représenter la quantité d'interactions impliquant un nœud est de mesurer les chaînes d'infection sortantes (voir partie 1.2.3.1), comme dans [Büttner et al., 2013]. Ainsi, l'ordre chronologique des interactions est pris en compte pour évaluer l'importance des nœuds.

Néanmoins, on peut arguer qu'il ne suffit pas d'échanger beaucoup pour être important. Connecter deux parties autrement isolées du réseau peut également être favorable à la propagation des maladies, indépendamment du degré de ce nœud jouant le rôle de pont. C'est pourquoi la mesure de la centralité d'intermédiarité (définition 1.8) est souvent utilisée pour cibler les nœuds à supprimer [Chen et al., 2008], pour stopper les phénomènes de diffusion. Sur un réseau dynamique, on peut préférer l'utilisation d'une des propositions de définition de centralité d'intermédiarité temporelle (voir partie 1.2.3.2).

En supprimant les nœuds les plus importants, quelle que soit la notion d'importance considérée (degré, centralité d'intermédiairité, etc.), le but recherché est de déconnecter le réseau. Au lieu de se baser sur le calcul des propriétés des nœuds pour évaluer leur importance, [Chen et al., 2008] proposent de rechercher directement l'ensemble des nœuds de taille minimale permettant de déconnecter le réseau. Ainsi, les auteurs s'inspirent de l'algorithme *nested dissection* [Lipton et al., 1979], pour identifier l'ensemble des nœuds à vacciner de taille minimale. Cet algorithme a pour but de diviser le réseau en groupes de nœuds de même taille, fixée en paramètre. Ce résultat doit être obtenu en supprimant le plus faible nombre de nœuds possible. Autrement dit, l'algorithme minimise la fraction des nœuds à retirer pour atteindre le résultat souhaité. Les nœuds supprimés représentent la part de la population vaccinée par la stratégie de lutte. Les auteurs montrent que le nombre de nœuds à vacciner avec cette méthode est 5 à 50% plus faible que la vaccination des nœuds de plus fort degré, pour une diffusion de type SIR.

Dans certaines situations, une connaissance exhaustive des interactions n'est pas possible. Par exemple, lorsque les contacts au sein de populations humaines sont étudiées, il n'est pas toujours possible techniquement d'équiper les individus de capteurs, comme dans [Fournet and Barrat, 2014]. Les individus doivent alors rapporter leurs interactions, ce qui peut être source d'imprécisions, notamment en cas d'oubli. C'est dans ce cadre qu'a été développée l'*acquaintance immunization* [Cohen et al., 2003]. Tout d'abord, cette méthode sélectionne aléatoirement ensemble de nœuds. Puis, un nombre donné de leurs re-

lations (*i.e.* de leurs nœuds voisins) sont vaccinés aléatoirement. De cette façon, les nœuds de forts degrés ont une plus grande chance d'être choisis et vaccinés parmi les voisins des nœuds, alors qu'ils sont justement des acteurs importants de la diffusion.

Dans [Salathé and Jones, 2010], les auteurs proposent une stratégie pour cibler les nœuds jouant le rôle de pont entre différentes parties du réseau, dans le cas où seule une connaissance locale du graphe est possible. Les auteurs choisissent aléatoirement des nœuds de départ pour les diffusions. Pour chaque nœud rencontré sur une propagation, ils vérifient ses connexions : si le nœud n'est pas connecté à d'autres nœuds de la cascade, il est identifié comme important. En effet, un tel nœud est probablement connecté à une partie inexplorée du réseau, et jouerait donc le rôle de pont entre cette partie inexplorée et celle explorée par la cascade. Les auteurs trouvent que la taille moyenne des propagations, simulées par un modèle SIR, est plus faible lorsque les nœuds sont supprimés selon cette stratégie d'identification qu'avec la méthode *acquaintance immunization*.

Par ailleurs, le réseau construit à partir de la BDNI présente la particularité d'être constitué de différents types de nœuds, à savoir les élevages, les marchés, et les centres de rassemblement. Il est donc possible de prendre en compte leur type pour cibler les nœuds et les liens à supprimer. Par exemple dans [Rautureau, 2012], la totalité des centres et des marchés est enlevée des séquences de réseaux annuels, mensuels ou hebdomadaires, et les tailles des GSCC sont mesurées. Cette méthode de suppression est efficace pour les réseaux mensuels et hebdomadaires, mais ne parvient pas à faire chuter la taille des GSCC pour les réseaux annuels. L'auteure note cependant la contrainte importante qu'aurait la mise en place d'un tel dispositif sur le commerce, en termes de limitation des échanges. Elle souligne la nécessité d'utiliser des mesures de centralité pour affiner le ciblage des nœuds et diminuer le nombre de suppressions. [Kao et al., 2006] s'intéresse quant à eux à la suppression de liens selon leur type, c'est-à-dire selon les types du nœud d'origine et du nœud de destination. Dans cette étude, les auteurs montrent que les liens de type élevage vers marché sont importants dans le réseaux d'échanges bovins de Grande-Bretagne. Ils sont donc supprimés donc lorsqu'ils permettent de relier deux marchés. Autrement dit, si un élevage achetant un animal à un premier marché en vend un à un autre marché, alors cette dernière interaction est interdite, et supprimée du réseau. Ils observent une chute de la taille de la plus grande composante connexe, ce qui tend à confirmer l'importance qu'a ce type de lien pour structurer le réseau et participer à la diffusion.

Dans le contexte de la propagation de maladies animales, les mesures de luttes peuvent être appliquées spécifiquement à des zones géographiques. C'est notamment le cas de l'abattage préventif, voir l'exemple de la grippe aviaire en 2017 [1]. C'est pourquoi [Büttner et al.,

1. Legifrance, arrêté du 5 janvier 2017 définissant les zones géographiques dans lesquelles un abattage préventif est ordonné (...) pour la maîtrise de l'épizootie d'influenza aviaire due au virus H5N8 dans certains départements.https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000033793904&categorieLien=id

2016] modélisent une stratégie de suppression basée sur la distance géographique, pour représenter les mesures de lutte telles que les restrictions commerciales, la vaccination sélective ou l'abattage préventif, qui peuvent être mis en place dans les zones entourant les exploitations infectées. La propagation d'une maladie est simulée selon un modèle SIR. Lorsqu'un nœud est infecté, la distance euclidienne le séparant de ses voisins est mesurée. Si cette distance est inférieure à une valeur donnée en paramètre (de 1 à 10 km), les voisins concernés sont supprimés du réseau, et ne peuvent donc pas être contaminés. Les auteurs comparent cette stratégie avec des mesures de luttes basées sur l'importance des nœuds. Ils montrent que cette méthode est moins efficace que celles basées sur les centralités des nœuds, ou les chaînes d'infection.

Dans [Le Menach et al., 2006], les auteurs modélisent la diffusion de la grippe aviaire A/H5N1 avec un modèle à cinq compartiments (susceptible, latent, infecté mais non détecté, infecté et détecté par les autorités, et abattu). Ils comparent deux stratégies de lutte : l'abattage des volailles des exploitations dans une aire autour d'un nœud infecté ; et la détection rapide de la maladie suivie de l'abattage ciblé de l'exploitation contaminée. Ils montrent que la dernière stratégie est la plus efficace. L'amélioration de la détection précoce et la mise en place rapide d'une mesure de lutte sont donc les éléments clés permettant une lutte efficace contre la propagation de cette maladie.

Lorsque des nœuds sont supprimés, avec eux disparaissent les liens qui leurs sont associés. Ceci a donc un impact sur les voisins de ces nœuds, qui voient leur nombre d'interaction diminuer. Il est possible que ces voisins cherchent à obtenir des animaux d'autres exploitations, pour compenser la perte des échanges avec les nœuds supprimés. Dans [Belik et al., 2015], dès qu'un nœud est détecté comme infecté (t instants après son infection), il est mis en quarantaine (*i.e.* supprimé du réseau). Les exploitations, avec qui le nœud supprimé interagissait, reçoivent leurs animaux d'autres nœuds tirés aléatoirement parmi les sains et les infectés non détectés. [Belik et al., 2015] mesurent alors la prévalence [2] au cours du temps, comme mesure de l'efficacité de leur méthode. Autrement dit, si la stratégie est efficace, elle doit réduire la prévalence. Ils étudient l'impact de la durée du temps de détection sur la prévalence, en faisant varier la durée de l'épidémie, c'est-à-dire sa période infectieuse. Ils montrent que les maladies avec période infectieuse inférieure à 3 mois et un temps de détection d'une semaine sont efficacement contrôlés avec cette stratégie de lutte, prenant en compte la réaffectation des interactions des exploitations saines.

4.1.2 Analyses rétrospectives et prédictives

Les données servent à établir les classements des nœuds selon leurs propriétés (par exemple, leur degré). Ceux-ci serviront à choisir les éléments clés à supprimer. Lorsque

2. Prévalence : nombre d'infectés total à un instant donné

ces mêmes données sont également utilisées pour la simulation des propagations (voir par exemple [Rautureau, 2012]), on parle d'analyse rétrospective : si les données avaient été accessibles au moment de la crise sanitaire, l'analyse montre quelle aurait été l'efficacité de la stratégie d'identification choisie.

Si les analyses rétrospectives sont de grande importance pour comprendre quels éléments du réseau ont joué un rôle important dans la diffusion, tester l'efficacité des stratégies de suppression au cours d'analyses prédictives est la première étape vers une utilisation de ces mesures dans des situations réelles de crises sanitaires. En effet, il est raisonnable de penser qu'un certain délai existera entre les dates effectives des contacts entre nœuds, le calcul de l'importance des éléments d'un réseau et la mise en œuvre de la mesure. Il faut donc que les méthodes d'identification ne soient pas trop sensibles aux fluctuations d'une période à une autre de l'importance des nœuds et des liens dans les données, observées au chapitre 2. Autrement dit, les nœuds et liens identifiés comme importants à un moment donné doivent l'être toujours suffisamment dans le futur pour permettre une réduction des tailles de propagation efficace. Ainsi, des études comme par exemple [Bajardi et al., 2011] identifient des nœuds à supprimer dans une période de temps fixée et selon les critères choisis, et simulent la suppression sur des données ultérieures à cette période.

Sur l'exemple de la BDNI, [Dutta et al., 2014] compare le calcul des centralités (degré, centralité d'intermédiarité, centralité de proximité, partie 1.2.3) sur deux types de données passées :

— les données précédant immédiatement les données étudiées ;
— les données correspondant à une période de temps similaire dans le passé, par exemple, le 1*er* trimestre de l'année précédente si l'on étudie le 1*er* trimestre de l'année en cours.

Ce dernier protocole part de l'hypothèse que les acteurs ont des comportements similaires d'année en année aux mêmes périodes de temps (même comportement par exemple chaque premier trimestre). Il tire donc parti de l'observation de cycles saisonniers dans les données d'échanges d'animaux, comme nous avons pu le voir au chapitre 2.3.1. Ainsi, il serait plus efficace d'utiliser des informations concernant une même période de temps, d'une année ultérieure. Cependant, [Dutta et al., 2014] montre que le protocole utilisant les données passées les plus récentes est plus efficace que celui utilisant des données d'une période passée équivalente. Selon cette étude, il vaut mieux utiliser les données les plus récentes à disposition pour mettre en place une stratégie de suppression. Toutefois, les tailles des GSCC sont réduites plutôt efficacement, quel que soit le protocole choisi. La validité des classements de centralité reste donc bonne au cours du temps, c'est-à-dire, les nœuds identifiés comme les plus importants du réseau dans le passé le sont globalement toujours autant par la suite.

[Bajardi et al., 2011] arrivent à une conclusion différente dans leur étude. Sur les données du réseau d'échanges d'animaux en Italie, ils montrent que le calcul du degré des

nœuds d'un réseau mensuel et l'utilisation du classement obtenu pour supprimer des nœuds dans le réseau mensuel suivant dans le temps ne permet pas de réduire efficacement la taille de la plus grande composante connexe (ne prenant pas en compte l'orientation des liens). Selon cette étude, la validité au cours du temps des classements est plutôt mauvaise. Les stratégies classiques d'identification des nœuds importants ne seraient donc pas adaptées.

4.2 Notre méthode

Notre objectif est d'étudier l'efficacité de la prise en compte de l'information temporelle sur les échanges pour l'identification des éléments clés du réseau dans les processus de diffusion. Pour ce faire, nous proposons dans cette section des méthodes de suppression basées sur les cascades simulées. Nous comparons ensuite ces stratégies à celles basées sur la suppression des nœuds de plus forts degrés (définition 1.6), de plus fortes activités (définition 1.7), et de plus fortes centralités d'intermédiarité (définition 1.8). Nous les comparons également à la suppression des liens selon leurs poids (définition 1.7). Notons que le calcul de la centralité d'intermédiarité sera approximé en tirant aléatoirement un cinquième des nœuds comme pivots [3]. Cette stratégie de choix des pivots est en effet recommandée par [Brandes and Pich, 2007].

Nous avons à disposition des méthodes de suppression ciblant différents éléments (nœuds ou liens). Il faut donc pouvoir les comparer selon des volumes de suppression équivalents. Cette partie s'achève donc avec la présentation d'une méthode simple de comparaison.

4.2.1 Occurrences dans les cascades

Dans le chapitre précédent, nous avons calculé les tailles des cascades des nœuds (définition 1.5), qui représentent l'impact potentiel de chaque cascade sur le réseau. On peut se demander à quel point ces cascades partagent des nœuds et des liens, et comment nous pourrions utiliser cette information pour identifier les éléments favorisant la propagation. Autrement dit, la suppression pourrait cibler les nœuds et liens partagés par de nombreuses cascades, étant a priori des passages privilégiés pour la diffusion des maladies. La notion d'importance sous-jacente dépend alors de la fréquence d'apparition dans les cascades. Nous reprenons donc la notion de cascade (voir partie 1.2.3.1) pour y compter le nombre de fois N_{occ} que chaque nœud, lien statique (couple de nœuds), et lien temporel y apparaissent. Avec $\mathcal{W}$ l'ensemble des nœuds temporels (t, v) de départ des diffusions, respectant

3. Un pivot est un nœud utilisé comme point de départ d'un *BFS* (et donc d'une cascade) sur un graphe, voir partie 1.2.3.1

les conditions définies en 3.2.2, et d la durée de propagation fixée :

$$\begin{aligned} N_{occ}(v) &= \sum_{(t_u,u)\in\mathcal{W}} |\{(x,y,z) \in \mathscr{C}_d(t_u,u) \ tq \ y = v \ ou \ z = v\}| \\ N_{occ}(u,v) &= \sum_{(t_w,w)\in\mathcal{W}} |\{(x,y,z) \in \mathscr{C}_d(t_w,w) \ tq \ y = u, \ z = v\}| \\ N_{occ}(t,u,v) &= \sum_{(t_w,w)\in\mathcal{W}} |\{(x,y,z) \in \mathscr{C}_d(t_w,w) \ tq \ x = t, \ y = u, \ z = v\}| \\ &= |\{(t_w,w) \ tq \ (t,u,v) \in \mathscr{C}_d(t_w,w)\}| \end{aligned} \tag{4.1}$$

Ainsi, ces mesures prennent en compte l'information temporelle contenue dans les chemins de propagation pour évaluer l'importance des nœuds et liens rencontrés.

L'occurrence de chacun de ces éléments est dépendant du nombre de cascades calculées. En effet, selon notre protocole présenté en partie 3.2.2, le nombre de cascades simulées sur une année est égal au nombre de nœuds actifs durant l'année considérée pour lesquels il existe un temps de départ respectant les conditions décrites, soit $|\mathcal{W}|$ cascades. Sachant qu'un lien temporel (t,u,v) ne peut apparaître par définition qu'une seule fois par cascade, et que les temps de départ tirés sont distribués uniformément, $\frac{|\mathcal{W}|}{k}$ (avec k le nombre de *snapshots*) est la valeur maximale que $N_{occ}(t,u,v)$ peut obtenir théoriquement. Par exemple en 2015, pour des propagations d'un mois, $|\mathcal{W}| = 165\ 829$. Un lien temporel peut donc théoriquement apparaître dans au plus $\frac{|\mathcal{W}|}{12}$, soit 13 819 au maximum. Et en effet, on observe que les liens temporels ayant le plus grand nombre d'occurrences dans les cascades apparaissent environ 12 800 fois en 2015, soit environ 93% du score maximal théorique. Cette observation tendrait à montrer que les liens temporels les mieux classés apparaissent dans la grande majorité des cascades d'une durée d'un mois. On peut donc espérer ralentir significativement les propagations sur cette période en supprimant ces liens.

N_{occ} étant dépendant du nombre de cascades calculées, le nombre d'éléments classés (nœuds ou liens), et par suite le nombre de suppressions pouvant être effectuées, est limité par le nombre d'éléments distincts rencontrés lors du calcul des cascades. Ainsi en 2015, on ne pourra pas supprimer plus de 673 994 (26%) liens temporels, 410 152 (40%) liens statiques, et 172 090 (97%) nœuds. Toutefois, de tels pourcentages de suppression seraient extrêmement contraignant pour le réseau, en interrompant une part importante des échanges. Si de tels volumes de suppression devaient être utilisés pour stopper la propagation de maladies, on pourrait considérer que ces méthodes ne parviennent pas à identifier les nœuds et liens importants pour la diffusion.

4.2.2 Volume de suppressions : base de comparaison

Dans la littérature, la taille des propagations est généralement exprimée en fonction du nombre de nœuds supprimés, par exemple dans [Bajardi et al., 2011, Dutta et al., 2014]. Si l'on veut comparer des méthodes de suppression de nœuds et de liens comme c'est le cas ici, il faut donc trouver une base de comparaison commune.

Lorsque l'on supprime un nœud du réseau, on retire également tous ses liens avec les autres nœuds, soit autant de liens que le degré du nœud. De même, retirer les liens statiques entre deux nœuds peut revenir à supprimer dans les données un nombre très important de liens temporels. Par exemple, retirer le nœud de plus fort degré entrant (obtenant un score de 1583), revient à retirer 5279 liens du réseau statique correspondant et 16066 liens temporels des données. C'est pourquoi, nous convertissons le nombre de nœuds et le nombre de liens statiques supprimés en leur nombre de liens temporels correspondant, que nous notons $\mathscr{V}$ pour volume : pour un nœud, on compte dans les données à combien de liens temporels il participe ; de même, pour un lien statique, on compte le nombre de fois qu'il apparaît dans les données. Ceci étend au contexte temporel l'approche utilisée par exemple par [Magnien et al., 2011] pour convertir un nombre de nœuds supprimés en un nombre de liens (statiques). En résumé, si l'on considère le flot de liens $L = (T, V, E)$, on définit :

$$\begin{aligned} &\textit{le volume de suppression d'un noeud } u \in V \textit{ par } \mathscr{V}_u = |\{(t,u,v) \in E\} \cup \{(t,v,u) \in E\}| \\ &\textit{le volume de suppression d'un lien statique } u,v \in V \textit{ par } \mathscr{V}_{u,v} = |\{(t,u,v) \in E\}| \\ &\textit{le volume de suppression d'un lien temporel } (t,u,v) \in E \textit{ par } \mathscr{V}_{t,u,v} = 1 \end{aligned} \tag{4.2}$$

Ainsi, les volumes de suppressions sont comparables, quelles que soient les stratégies considérées. Dans ce travail, le terme "coût" se réfère à ce volume de suppression.

Dans la littérature, un pourcentage faible de nœuds est généralement supprimé : par exemple [Dutta et al., 2014] suppriment jusqu'à 4% de nœuds pour stopper les propagations ; [Rautureau, 2012] constate qu'il suffit de supprimer environ 1% des nœuds, selon le protocole, pour faire disparaître la GSCC. Le tableau 4.1 présente les pourcentages de liens supprimés correspondants dans l'année 2015, pour trois stratégies de ciblage des nœuds (voir en annexes les autres stratégies). On constate que le nombre de nœuds supprimés équivaut en termes de volume de suppression à des fractions très élevées du réseau : plus de 45% des interactions ne pourraient avoir lieu dès que $0,1\%$ des nœuds sont supprimés. Les échanges d'animaux sont donc grandement impactés par des pourcentages de suppressions de nœuds qui semblent faibles au premier abord. Le choix de la mesure de l'efficacité de la suppression est donc un paramètre important du protocole de comparaison, car il conditionne la vision de l'acceptabilité d'une stratégie. Autrement dit, une même stratégie peut sembler inacceptable par rapport à la contrainte exercée sur les échanges du réseau,

TABLE 4.1 – Pourcentages de liens temporels supprimés en fonction du pourcentage de nœuds, pour l'année 2015, pour trois exemples de stratégies de suppression (degré sortant k_{out}, activité sortante $\mathcal{A}$, et centralité d'intermédiarité BC).

% des nœuds	0,1	0,2	0,4	0,8	1
k_{out}	46	61	74	79	80
$\mathcal{A}_{out}$	47	62	75	80	80
BC	46	61	73	79	80

alors qu'évaluée par une autre mesure, elle apparaîtrait peu contraignante en termes de pourcentages de suppression.

Nous proposons des stratégies d'identification des nœuds, liens statiques, et liens temporels importants pour la diffusion. Elles consistent à évaluer le nombre de fois que ces éléments apparaissent dans des cascades de durée fixée. Comparativement à d'autres stratégies comme le degré, elles prennent en compte l'information temporelle sur les interactions.
Les stratégies d'identification n'aboutissant pas à la suppression du même type d'élément dans le réseau (nœuds ou liens), nous proposons de quantifier les ensembles de nœuds et de liens statiques par leur nombre de liens temporels correspondant.

4.3 Résultats expérimentaux

Nous utilisons un protocole similaire à celui du chapitre 3 pour calculer les occurrences des nœuds et liens dans les cascades bornées dans le temps. Puis nous calculons les tailles des cascades (*i.e.* leurs chaînes d'infection sortantes, définition 1.5) en fonction du nombre de suppressions effectuées, avec des temps de départ pour les cascades correspondantes différents de ceux utilisés pour le calcul des classements. Le but est de déterminer la pertinence effective des classements. Pour une meilleure lisibilité des résultats, nous nous concentrons ici sur la présentation du pire cas de propagation, c'est-à-dire la comparaison des tailles des plus grandes cascades.

L'efficacité des stratégies de suppression est testée dans deux cas de figure. Tout d'abord, nous nous positionnons dans le cas d'une analyse rétrospective. Nous recherchons quelles stratégies auraient été les plus efficaces à une période de temps donnée. Ensuite, nous étudions la validité au cours du temps des classements, et menons donc une analyse prédictive.

Ce type d'analyse est un premier pas vers une utilisation en conditions réelles des stratégies d'identification des éléments clés d'un réseau. Nous concluons alors quant à l'efficacité de ces mesures de suppression prenant en compte l'information temporelle.

4.3.1 Comparaison des classements des méthodes

Avant d'étudier l'efficacité des mesures de suppression pour réduire les tailles des cascades, nous comparons les classements obtenus. Nous cherchons ici à évaluer à quel point ces mesures permettent d'identifier des nœuds et des liens différents, résultats que nous mettrons en perspective dans la partie suivante.

Les stratégies de suppression que nous avons implémentées ciblent des éléments de natures différentes : nœuds, liens statiques ou liens temporels. Pour les comparer, nous procédons d'une manière similaire à la partie 4.2.2 : nous les transformons en ensembles d'éléments similaires. Nous choisissons de ramener chaque ensemble d'éléments à supprimer A' en leur équivalent en liens temporels A :

— Soit A' un ensemble de nœuds, $A = \{(t, u, v) \in E, u \in A'\} \cup \{(t, v, u), u \in A'\}$

— Soit A' un ensemble de liens statiques, $A = \{(t, u, v) \in E, (u, v) \in A'\}$

Ainsi, on obtient des ensembles de liens temporels, que l'on peut comparer au moyen du calcul de l'indice de Jaccard, noté J. Soient A et B des ensembles de liens temporels :

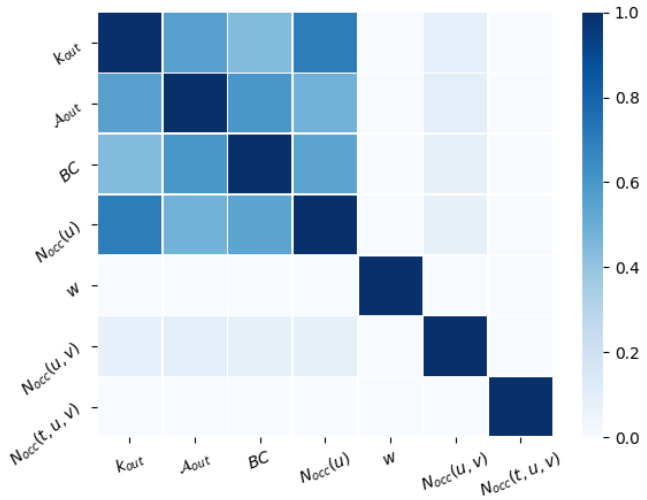
$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{4.3}$$

Si la valeur obtenue par le calcul de l'indice de Jaccard est proche de 1, les classements comparés sont similaires. Au contraire, un résultat proche de 0 indique que les classements sont très différents. En pratique, les classements A et B ne sont pas toujours strictement de même taille, notamment concernant les classements de nœuds. En effet, les nœuds les plus centraux ont souvent un nombre très élevé d'interactions. Les volumes de liens supprimés augmentent donc brusquement lorsque l'on supprime un nœud supplémentaire. Il est donc difficile d'obtenir des volumes identiques, notamment pour de faibles nombres de suppressions.

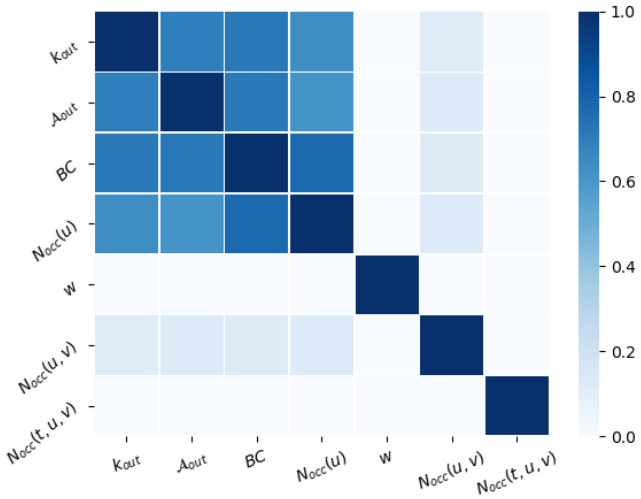
La figure 4.2 présente, pour différents volumes de suppression, les résultats d'indice de Jaccard, permettant de comparer les stratégies. Lorsque 10 nœuds sont supprimés, ce qui correspond à environ 8% de liens temporels (figure 4.2a), on observe un cluster de scores d'indice de Jaccard pour les classements de nœuds :

— le degré et l'occurrence des nœuds produisent les classements les plus similaires ($J \simeq 0,8$) ;

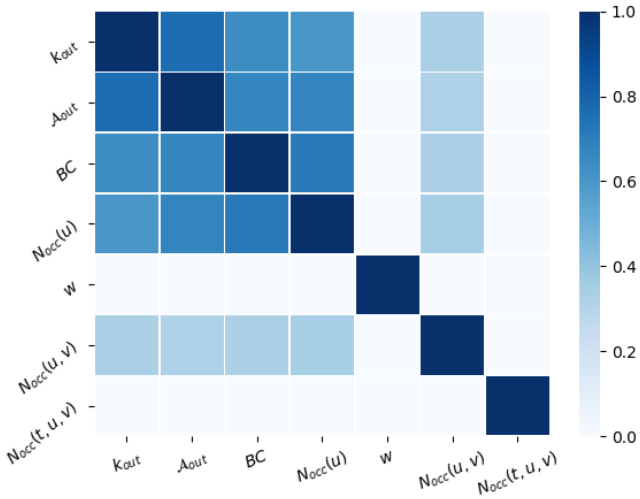
— avec des valeurs d'indice de $0,6$, le degré et l'activité, la centralité d'intermédiarité et le degré, et enfin la centralité d'intermédiarité et l'occurrence des nœuds produisent



(a) 8% de liens temporels supprimés



(b) 15% de liens temporels supprimés



(c) 33% de liens temporels supprimés

FIGURE 4.2 – Indice de Jaccard calculé pour chaque couple de stratégies de suppression, calculées sur 2015. On note k_{out} le degré sortant, $\mathcal{A}_{out}$ l'activité sortante, BC la centralité d'intermédiarité, $N_{occ}(u)$ le nombre d'occurrences des nœuds, w le poids des liens statiques, $N_{occ}(u,v)$ le nombre d'occurrences des liens statiques, et $N_{occ}(t,u,v)$ le nombre d'occurrences des liens temporels dans les cascades.

des classements relativement similaires ;

— l'activité et l'occurrence des nœuds, et la centralité d'intermédiarité et le degré présentent ensuite des classement légèrement moins similaires, avec $J \simeq 0,4$.

Concernant les classements de liens :

- les classements de liens sont peu similaires entre eux ($J < 0,2$) ;
- les classements des nœuds comparés aux classements des liens sont peu similaires ($J < 0,2$).

Lorsque l'on supprime 25 nœuds (figure 4.2b), soit environ 15% des liens temporels :

- la similarité entre classements de nœuds croît encore ($J \geq 0,6$) ;
- la similarité entre classements de liens, ainsi que la similarité entre classements de nœuds et de liens demeure faible ($J \leq 0,2$)

De même, une suppression d'environ 33% liens temporels, soit une centaine de nœuds, aboutit (figure 4.2c) :

- à de fortes similitudes entre classements de nœuds ;
- à une similarité entre classements de liens qui reste faible ($J \leq 0,2$) ;
- à une similarité entre classements de liens et de nœuds qui reste faible ($J \leq 0,2$), à l'exception du classement par score $N_{occ}(u,v)$. Cette stratégie est la seule parmi les méthodes de suppression de liens à voir ses scores d'indice de Jaccard augmenter avec l'augmentation des pourcentages de suppression : elle atteint des scores de $0,4$ lorsque comparée aux méthodes de suppression de nœuds.

Plus les classements sont similaires, plus on peut s'attendre à ce que les résultats de suppression le soient également. Ainsi, on peut s'attendre à ce que l'occurrence des nœuds donne des résultats de suppression qualitativement proches de ceux obtenus avec le degré. Au contraire, on s'attend à ce que les suppressions de liens statiques et temporels selon leur occurrence dans les cascades donnent des résultats très différents des autres classements.

Cette étude nous permet de conclure qu'intégrer l'information temporelle via une mesure d'activité, ou par une mesure du nombre d'occurrences dans les cascades, ne mène pas à supprimer des nœuds différents par rapport aux mesures de centralités classiques. Par contre, mesurer le nombre d'occurrences des liens dans les cascades permet de cibler des liens très différents de ceux qui pourraient être sélectionnés par l'étude de leur poids w dans les données.

> Les classements de nœuds, utilisant l'information temporelle ou non pour leur construction, sont très similaires. On peut donc supposer qu'ils auront des performances semblables pour diminuer les tailles de propagations.
> Au contraire, les classements de liens sont très différents entre eux et avec les classements de nœuds. Ils entraînent donc la suppression d'éléments différents, et auront donc potentiellement des performances très différentes.

4.3.2 Analyse rétrospective

Afin de comparer des volumes équivalents, les suppressions sont exprimées en leur équivalent en liens temporels (définition 4.2), que l'on rapporte ensuite au nombre total de liens temporels du réseau. Le tableau de comparaison des pourcentages et nombres de suppressions correspondant est à retrouver en annexes.

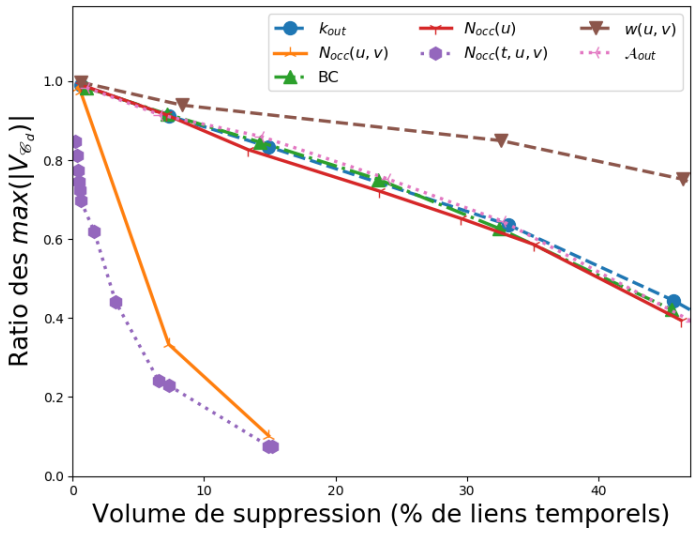


FIGURE 4.3 – Ratio des tailles de cascades maximales de 2015 en fonction du nombre de suppressions, exprimé volume de suppression, pour des cascades de durée d'un mois.

On mesure pour chaque stratégie, et pour chaque volume de suppression, le ratio de la taille de la plus grande cascade avec suppression, sur celle sans suppression. Autrement dit, on calcule $max(|V_{\mathscr{C}_d(t,u)}|)_{(t,u)\in\mathcal{W}}$ sur le réseau avec, puis sans suppressions, et on divise ces deux valeurs. Les résultats sont comparés en figure 4.3. On observe trois ensembles de courbes de résultats :

— le premier est constitué des stratégies de ciblage des nœuds, qui sont qualitativement similaires, comme attendu ;
— le ciblage selon les scores $N_{occ}(t,u,v)$ et $N_{occ}(u,v)$ forment un ensemble de courbes qualitativement similaires ;
— le dernier ensemble est constitué d'une seule stratégie, la suppression selon le poids des liens statiques dans les données.

La suppression selon le poids des liens statiques dans les données est la méthode permettant une diminution du pire cas de propagation la plus faible : par exemple pour une suppression d'environ 8% de liens temporels, la taille maximale diminue de 6%.

Les suppressions selon le degré (entrant, sortant et total) et l'activité (entrante, sortante, totale) des nœuds donnent des résultats qualitativement similaires ; nous ne représentons ici que le degré et l'activité sortants, qui sont représentatifs des autres méthodes. En supprimant par exemple environ 8% des liens temporels, la taille maximale potentielle de propagation diminue de 9% environ. Or, les mesures d'activités intègrent l'information temporelle sur le volumes d'échanges des noeud, pour établir leurs classements. Cette utilisation de l'information temporelle ne permet pas ici d'améliorer les résultats. De même, la mesure du N_{occ} des nœuds ne permet pas non plus de réduire plus fortement des tailles de propagation que les mesures basées sur le degré. Cette similarité dans les résultats peut s'expliquer par les classements de nœuds similaires obtenus par ces méthodes.

Supprimer les nœuds selon leur centralité d'intermédiarité réduit les tailles des cascades de manière comparable aux autres stratégies de suppression de nœuds. Pourtant, cette stratégie est souvent considérée comme plus efficace, comme par exemple dans [Dutta et al., 2014]. Ceci peut potentiellement s'expliquer par un nombre de suppressions plus faible que dans la littérature : retirer 1 à 10% des nœuds comme cela y est fait correspond à empêcher respectivement 46 à 80% des échanges d'animaux d'avoir lieu, soit une fraction des interactions retirée exorbitante. Le comptage en volume de suppression changerait donc l'impression que l'on peut avoir sur l'efficacité relative de ses stratégies. Néanmoins, cette similarité des résultats n'est pas surprenante à la vue de la ressemblance du classement de la centralité d'intermédiarité avec ceux des autres méthodes de suppression de nœuds.

Finalement, les suppressions de liens temporels et statiques selon leur score N_{occ} sont les méthodes les plus performantes : par exemple pour environ 8% de liens temporels supprimés, ces méthodes permettent respectivement une baisse de 77 et 67% de la taille maximale potentielle de propagation.

Dans cette partie, nous avons comparé, rétrospectivement, l'efficacité des méthodes de suppressions de nœuds et liens intégrant une part de l'information temporelle, à des stratégies ignorant cette information.
Nous avons montré que toute information temporelle ne permet pas d'améliorer l'efficacité des stratégies d'identification des nœuds et liens clés de la propagation :

- prendre en compte le volume des interactions, via les mesures d'activité, n'est pas une stratégie de suppression temporelle plus efficace que son équivalent statique (calcul des degrés) ;
- prendre en compte l'information temporelle sur les cascades, via le comptage des nombres d'occurrences des nœuds, n'est pas non plus efficace ;

Au contraire, prendre en compte l'information temporelle sur les cascades, via le comptage des nombres d'occurrences des liens statiques ou temporels, améliore grandement l'efficacité de la suppression.

4.3.3 Analyse prédictive

Si l'on se projette dans une situation réelle d'utilisation des mesures d'importance des nœuds et des liens pour identifier les éléments à supprimer, il faudrait certainement prendre en compte l'existence d'un délai entre la notification par un éleveur d'un échange d'animal, l'enregistrement effectif du changement d'exploitation dans la BDNI et l'intervention en elle-même. Ainsi, à un instant donné, les interactions les plus récentes dans le réseau ne seraient pas connues et utilisables pour le calcul des centralités. L'identification des éléments clés pour la diffusion ne pourrait pas être réalisée en temps réel. Se pose alors la question de la sensibilité des stratégies d'identification à cette situation : les tailles des cascades sont-elles efficacement réduites malgré la prise en compte des interactions passées pour la détection des éléments clés du réseau dans le futur ?

Dans cette partie, nous supposons que les données de l'année $i-1$ peuvent être utilisées pour le calcul de l'importance, pour une mise en œuvre des stratégies de suppression sur la i-ème année. Nous nous concentrons ici sur l'exemple des années 2014 et 2015. L'année 2014 est donc transformée en graphe pour calculer des centralités en contexte statique : y sont calculés le degré sortant et la centralité d'intermédiarité. L'année 2014 est également utilisée sous forme de flot de liens pour calculer le score d'occurrence des liens statiques dans les cascades ($N_{occ}(u,v),\ \forall u,v \in V$). Nous n'utilisons pas le calcul du nombre d'occurrences des liens temporels dans les cascades, malgré le fait que ce soit la stratégie la plus efficace : par définition un lien temporel figurant dans un classement en 2014 ne peut pas être observé parmi les liens de 2015. Il faudrait pour les cibler se tourner vers l'utilisation de méthodes de prédiction des interactions, ce qui dépasse le cadre de cette thèse.

Une fois les centralités calculées sur l'année 2014, nous calculons les tailles des cascades sur le flot de liens de 2015, avec un pourcentage variable de suppressions effectuées (figure 4.4). Le degré sortant et la centralité d'intermédiarité produisent de nouveau des résultats similaires, et demeurent moins efficaces que la suppression de liens statiques selon leur score N_{occ}. Concernant cette dernière méthode, on observe une stagnation des tailles maximales potentielles d'infections vers environ 40% de nœuds atteints pour plus de 15% de liens temporels supprimés. Cette stagnation pourrait s'expliquer par la suppression de liens se trouvant en bout de chemins de propagation : tous les liens rencontrés sur les cascades ont été classés, même ceux n'y apparaissant qu'un nombre limité de fois ; or, ces liens ne sont pas forcément utiles pour la propagation, et pourraient mener à des impasses.

Par ailleurs, nous attirons l'attention du lecteur sur le fait que nous exprimons les suppressions en termes de pourcentages de liens temporels, et non pas en pourcentages de nœuds. Ainsi, concernant le degré et la centralité d'intermédiarité, si 30% de suppressions ne permettent une diminution que de 30% environ des tailles des cascades maximales, cela correspond à supprimer une centaine de nœuds dans le réseau, soit un pourcentage infime

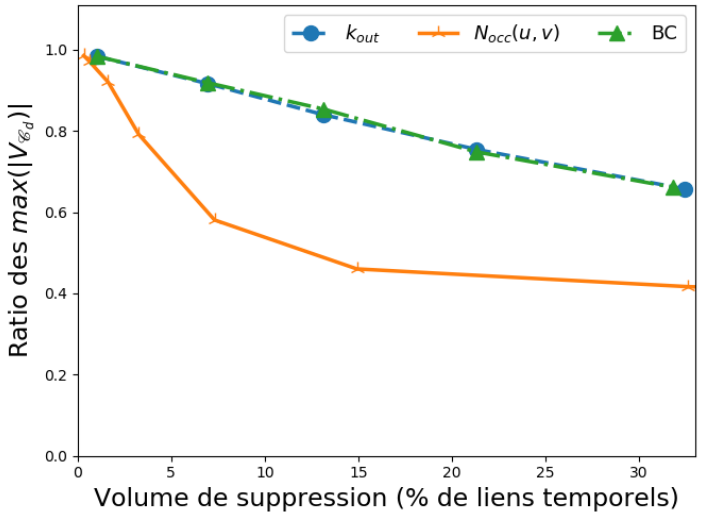


FIGURE 4.4 – Ratio des tailles de cascades maximales de 2015 de durée d'un mois, en fonction du nombre de suppression, exprimé en volume de suppression. Les classements de suppression sont obtenus sur 2014.

$(0,06\%)$. Si les suppressions étaient exprimées en termes de pourcentages de nœuds, la mesure des tailles des cascades révélerait donc, comme observé par [Dutta et al., 2014], que les suppressions selon les degrés et la centralité d'intermédiarité sont efficaces même lorsque l'on utilise des données passées pour construire les classements de nœuds.

Pour conclure, cibler les échanges d'animaux, c'est-à-dire supprimer les liens du réseau, est très efficace en dessous d'environ 15% des échanges sur l'année 2015. Au delà de ce seuil, supprimer plus de liens avec ces stratégies ne rend pas la lutte contre la propagation des maladies plus performante. Aussi, de nombreuses perspectives s'ouvrent pour continuer à améliorer les méthodes de suppression et à y intégrer la temporalité (voir chapitre 6).

Cette partie teste la capacité des stratégies à identifier les nœuds et les liens importants à un moment donné, et qui le demeurent dans les données futures. Utiliser l'information temporelle sur les chemins de propagation, via le calcul du nombre d'occurrences des liens statiques dans les cascades, permet une meilleure réduction des tailles des cascades que les deux stratégies statiques testées, à savoir le degré sortant et la centralité d'intermédiarité. Comme dans le cas d'une analyse rétrospective, la prise en compte de l'information temporelle semble donc essentielle pour prédire l'importance des éléments du réseau.

Conclusion

Dans cette partie, nous montrons que l'intégration d'une part de l'information temporelle du réseau, ici concernant les chemins de propagation ayant le plus de chances d'être empruntés en cas de crise, permet d'améliorer significativement les méthodes de suppression.

De plus, supprimer des liens, temporels ou statiques, permet le ciblage d'un nombre plus restreint d'éléments du réseau que les méthodes de suppression de nœuds. Ainsi, une mise en œuvre entraînerait probablement des coûts moindres.

Par ailleurs, exprimer les suppressions en termes de pourcentages de liens temporels, et non pas de pourcentages de nœuds, permet de se rendre compte que les méthodes de ciblage jugées efficaces dans la littérature entraînent la suppression d'un grand nombre d'interactions. La contrainte exercée sur le réseau par les stratégies de suppression est alors sous-évaluée. La métrique choisie a donc une grande influence sur notre vision de l'efficacité d'une stratégie.

Chapitre 5

Caractérisation de l'accessibilité des exploitations

Sommaire

Dans le chapitre 3, la mesure des cascades nous a permis d'obtenir la liste des interactions temporelles impliquées dans les propagations, sous la forme : date d'infection, nœud source, nœud infecté. Cette information nous a permis de compter le nombre d'infectés total, et de compter le nombre d'occurrences des nœuds et des liens dans les cascades pour proposer des stratégies d'identification de nœuds et liens importants dans le processus de diffusion (chapitre 4). Dans ce chapitre, nous étudions l'information sur la vitesse de propagation contenue dans les cascades, et son lien avec le nombre de nœuds atteints.

5.1 Diversité des vitesses de propagation

La première question à laquelle nous souhaitons répondre concerne la diversité de vitesses de propagation en fonction des tailles des cascades. Pour ce faire, nous cherchons

une mesure scalaire pouvant représenter cette vitesse de manière synthétique.

• **L'aire sous la courbe d'infection : un indicateur de la vitesse de propagation**

Des études précédentes de la diffusion sur des réseaux complexes laissent penser que les

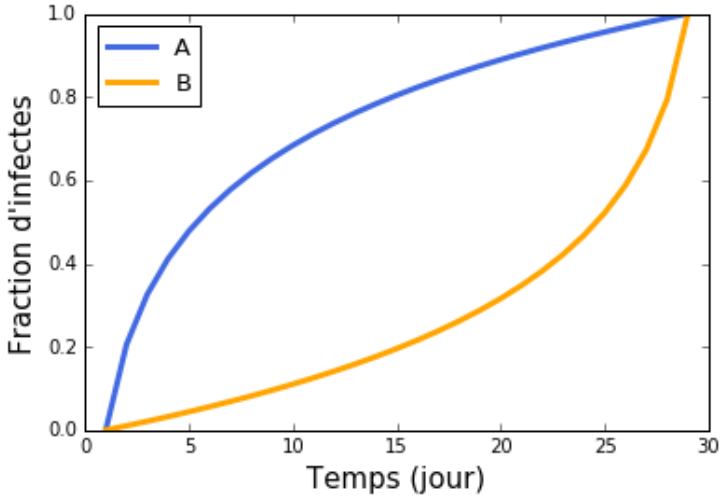


FIGURE 5.1 – Deux propagations théoriques : le scénario A représente une diffusion rapide puis saturant ; B représente une diffusion lente puis soumise à une accélération.

propagations ont une allure sigmoïde [Barthélemy et al., 2004]. Les auteurs de cette étude simulent un processus SI sur des données synthétiques, et observent un démarrage superlinéaire de la diffusion (modélisé par une exponentielle), puis une saturation : la croissance du nombre d'infectés ralentit puis devient nulle. La pente maximale, c'est-à-dire le nombre d'infecté par unité de temps de valeur maximale, peut alors être utilisée pour décrire la vitesse de la propagation. Cependant, ce calcul peut s'avérer non représentatif de la dangerosité de la diffusion, voir figure 5.1. En effet, cette figure présente deux scénarios de propagation, où les courbes sont de même pente maximale. Le scénario A représente une diffusion rapide au début de l'expérience, puis qui ralentit jusqu'à montrer un début de saturation. Dans le scénario B, la diffusion est lente au début de la période de temps étudiée, puis accélère brusquement. Avec le calcul de la pente maximale, les deux scénarios seraient jugés équivalents. Or, nous souhaitons pouvoir les différencier, notamment s'ils atteignent le même nombre de nœuds à la fin de la diffusion. La pente instantanée devrait a priori changer de manière importante au cours des scénarios A et B. Pour permettre la comparaison, il faudrait pouvoir ramener les valeurs de pentes instantanées à un score unique. Nous avons donc choisi de calculer l'aire sous la courbe du nombre d'infectés au cours du temps, abrégée *ASCI* :

$$ASCI_d(t_0, u) = \int_{t=t_0}^{t_0+d} |V_{\mathscr{C}_t(t_0,u)}| dt \tag{5.1}$$

avec d la durée des cascades et $|V_{\mathscr{C}_t(t_0,u)}|$ leur taille au temps t (définition 1.5). Nous obtenons ainsi une mesure cumulative du nombre d'infectés. Elle pourrait s'interpréter comme la pression infectieuse agrégée sur la période de propagation.

- **ASCI en fonction des tailles des cascades**

La figure 5.2 montre l'ASCI en fonction du nombre final d'infectés. Nous observons qu'il existe une certaine diversité d'ASCI, et donc de vitesses, pour un même nombre de nœuds atteints. Par exemple, pour 6000 infectés, l'ASCI est comprise entre 40000 et plus de 80000, soit deux fois plus. De plus, on remarque que le nuage de points est organisé en un faisceau de paraboles, et que son enveloppe a une forme que nous qualifierons d'aile par la suite. Dans la partie 5.3, nous chercherons à reconstruire ces deux caractéristiques, en modélisant la dynamique de propagation. En effet, modéliser les propagations permettrait d'identifier des paramètres simples expliquant la dynamique mise en œuvre.

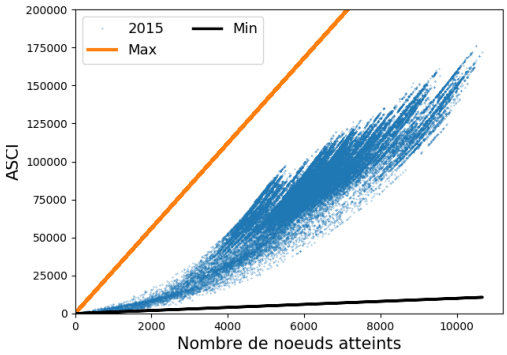


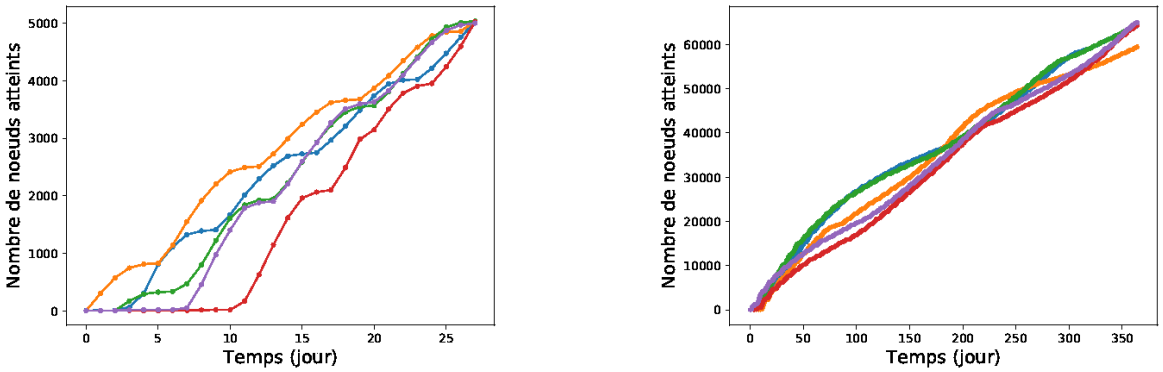
FIGURE 5.2 – Aire sous la courbe du nombre d'infectés au cours du temps, en fonction du nombre de nœuds atteints par les diffusions de 2015 de durée $d = 1$ mois. Les droites représentent les bornes théoriques minimale et maximale.

De plus, la figure 5.2 montre les bornes théoriques maximales et minimales d'ASCI qui seraient attendues :

— la borne maximale est obtenue lorsque le nombre final d'infectés est atteint au premier instant de temps (serait le cas d'un nœud de degré sortant égal au nombre final d'infectés), soit $ASCI = |V_{\mathscr{C}_d}| \cdot d$;

— la borne minimale théorique correspond au cas où le nombre final d'infectés est atteint au dernier temps de la propagation, soit $ASCI = d + |V_{\mathscr{C}_d}|$.

On observe que les valeurs obtenues lors des simulations sont très éloignées des bornes théoriques. Cela voudrait dire que tous les scénarios de cascades, intermédiaires à ceux représentés par les bornes, ne sont pas rencontrés dans les mesures.



(a) Cascades d'un mois (b) Cascades d'un an

FIGURE 5.3 – Allures typiques de propagations de différentes durées, choisies parmi des cascades de tailles similaires.

Il existe une grande variété de vitesses de propagation en fonction du nombre de nœuds atteints. Une relation parabolique relie ces deux variables, que nous cherchons à expliciter.
Les valeurs d'ASCI mesurées sont éloignées des bornes théoriques. Tous les scénarios de cascades n'existent donc pas dans les simulations.

5.2 Profils de propagation

Afin d'expliquer les écarts d'ASCI avec les bornes théoriques, on examine les profils de propagation, c'est-à-dire le nombre de nœuds atteints par une cascade au cours du temps. Le but est de chercher s'il existe un ou plusieurs profils typiques, afin de dégager les caractéristiques générales gouvernant leur dynamique. En effet, on peut supposer que l'existence d'un nombre restreint de profils différents induise que toutes les valeurs théoriques d'ASCI ne puissent pas être atteintes en pratique. En outre, nous voulons identifier des scénarios de propagation typique selon les tailles des cascades ou leur ASCI. Par exemple, les cascades de tailles maximales ont-elles un profil plus proche du scénario A de la figure 5.1 et celles de petite taille plus proche du scénario B ?

Tout d'abord, nous examinons un échantillon aléatoire de cascades d'un mois à un an, afin d'en étudier le profil et d'obtenir une première indication sur la diversité des profils existants. On représente quelques profils typiques en figure 5.3. On observe que les profils représentés sont globalement similaires, pour une durée donnée. Ils sont de plus assez éloignés des allures attendues en contexte épidémiologique, présentées en figure 5.1, ou

sigmoïdes. Certaines des cascades d'un mois (figure 5.3a) présentent une première phase de diffusion lente, suivie d'une phase nettement plus rapide, dont la vitesse semble affectée par les variations d'activités au cours des semaines (observées en partie 2.3.7). En effet, ces phases de croissances semblent se faire par paliers, de durées d'une semaine. Elles semblent néanmoins pouvoir être approximées de manière linéaire. Les autres cascades de la figure débutent directement par la phase linéaire, sans passer par un stade de diffusion lent. De même, les propagations d'un an (figure 5.3b) présentent une croissance linéaire, modulée par des phases lentes et des phases plus rapides, bien éloignée des profils théoriques attendus. Dans les deux cas, la durée de diffusion fixée ne permet pas d'atteindre de phase de saturation.

Il n'est pas exclu qu'une étude exhaustive des cascades mette en évidence d'autres types de profils que ceux observés ci-dessus. En effet, le nombre d'infectés final de la figure 5.3a sont similaires. Il est possible que le profil de propagation soit différent lorsque l'ordre de grandeur du nombre de nœuds atteints varie. C'est pourquoi, nous groupons à présent les cascades en fonction de leur taille, et étudions le profil moyen de chaque catégorie. Autrement dit, on calcule la fraction moyenne f des nœuds atteints au cours du temps, selon la définition suivante. Soient $L = (T, V, E)$ un flot de liens, d la durée des propagations, et $\mathcal{W}$ l'ensemble de leurs couples temps-nœud de départ. On pose I un intervalle d'entiers non négatifs. On définit la fraction moyenne du nombre de nœuds atteints en fonction du temps $f(t)$ comme :

$$f(t) = \frac{\sum_{(t,u)\in\mathcal{W}_I} \left(\frac{|V_{\mathscr{C}_t(t,u)}|}{|V_{\mathscr{C}_d(t,u)}|}\right)}{|\mathcal{W}_I|} \tag{5.2}$$

avec l'ensemble de nœuds temporels $\mathcal{W}_I = \{(t,u) \in \mathcal{W} \; tq \; |V_{\mathscr{C}_d(t,u)}| \in I\}$.

La figure 5.4 présente les résultats de $f(t)$ pour $t \in [0, d]$, lorsque les cascades de 2015 sont groupées en quartiles en fonction de leur taille. Pour rappel, la cascade de 2015 de taille maximale atteint 10 673 nœuds.

Le premier quartile contient 70% de cascades de tailles inférieures à 10 nœuds. La courbe correspondante a une forme mal définie et présente de fortes fluctuations du fait de la petite taille des cascades.

Concernant les courbes des trois autres quartiles, on remarque qu'elles se superposent. Ainsi, on n'observe pas de différence significative entre les profils moyens de propagation par quartile. Les cascades suivent donc des dynamiques de diffusion similaires, indépendamment de leur nombre final de nœuds atteints.

Dans la partie précédente, nous avons vu qu'à un même nombre d'infecté correspond une grande variété de valeurs d'ASCI, et donc de vitesses de propagation. On peut donc

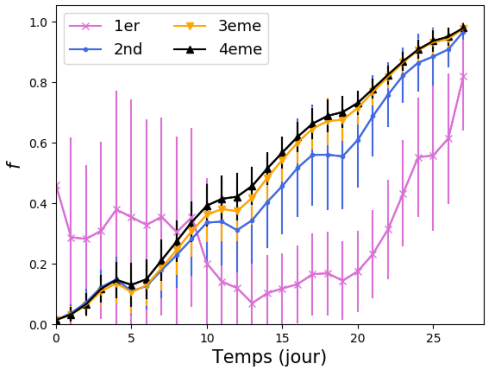


FIGURE 5.4 – Fraction moyenne de nœuds atteins par les diffusions de 2015 au cours du temps (cascades de durée d'un mois), en fonction du quartile d'appartenance de leur taille finale. On représente à chaque instant de temps l'intervalle de confiance.

supposer que cette distinction par quartile des tailles de propagation n'est pas une représentation adéquate pour étudier les différents profils. Il peut être préférable de distinguer les profils selon leur d'ASCI. Par exemple, on peut penser que les cascades obtenant les ASCI maximales pour une taille donnée aient un profil de propagation similaire au scénario A de la figure 5.1, alors que celles obtenant les scores minimaux correspondent au scénario B. Pour tester cette hypothèse, pour chaque valeur de taille de cascades, on enregistre les propagations d'ASCI maximales et minimales correspondantes. Puis, pour le groupe des cascades de forte ASCI et pour celui des faibles ASCI, on calcule la fraction moyenne du nombre de nœuds atteints au cours du temps (équation 5.2) sur l'ensemble des cascades d'ASCI maximums puis sur l'ensemble des cascades d'ASCI minimums. Nous sommes donc en train de produire le profil typique des cascades d'ASCI maximales et minimales.

La figure 5.5 montre les résultats : on n'observe pas de différence très significative dans les profils de propagation des diffusions d'ASCI minimales et maximales. Autrement dit, si la courbe moyenne maximale infecte plus rapidement de nouveaux nœuds, son profil est très similaire à celui de la courbe moyenne minimale. Les cascades semblent pouvoir être approximées de la même façon, quelle que soit leur ASCI.

Les cascades mesurées partagent certaines caractéristiques communes, indépendamment de leur taille et de leur ASCI. Leur dynamique semble donc pouvoir être modélisée de la même façon. L'examen d'exemples spécifiques implique qu'une approximation linéaire des profils serait adaptée, prenant en compte l'existence d'une phase de propagation négligeable, puis d'une phase rapide.
A noter que les profils de propagation observés sont bien sûr dépendants des paramètres fixés pour leur mesure, ici une durée de diffusion inférieure à 1 an.

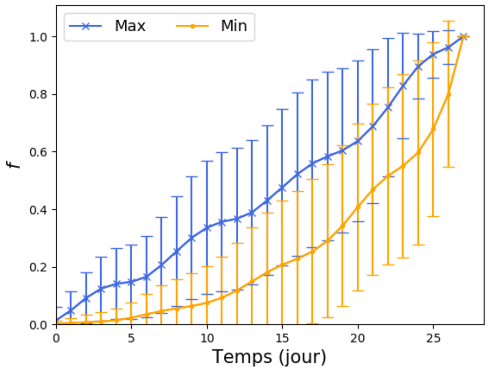


FIGURE 5.5 – Les cascades de 2015, bornées à un mois, sont divisées en groupes selon leur taille. Au sein d'un groupe, l'écart de taille entre la plus grande et la plus petite cascade est de 10 nœuds. Les ACSI sont comparées par groupe, pour ne retenir que celles d'ASCI maximales d'une part, et celles d'ASCI minimales d'autre part. La fraction moyenne de nœuds atteins en fonction du temps est calculée, pour les cascades obtenant les ASCI maximales puis pour celles obtenant les scores minimaux. L'intervalle de confiance est représenté à chaque instant de temps.

5.3 Modélisation de la dynamique de propagation

Modéliser la diffusion des maladies, et confronter les observations avec celles obtenues lors des simulations des chapitres précédents, permet de tester la validité de différentes hypothèses décrivant la dynamique de propagation.

5.3.1 Modèle à temps d'interactions aléatoires

Un modèle simple pour simuler la dynamique de propagation consiste à tirer aléatoirement des temps d'infection pour chacun des nœuds atteints. Compte-tenu des conclusions tirées de la partie précédente, ce modèle peut sembler trop élémentaire pour pouvoir reproduire les profils de propagation observés. Cependant, ce travail a été réalisé en même temps que la partie 5.2, et nous trouvons intéressant de montrer les résultats produits pour donner au lecteur une meilleure intuition de la relation entre les tailles et les ASCI des cascades. Pour chaque résultat de taille de cascade, obtenus dans le chapitre précédent, un même nombre de temps d'infection est tiré. L'ASCI est ensuite calculée sur la propagation ainsi modélisée :

— Soit l'ensemble des tailles de cascades mesurées $I = \{|V_{\mathscr{C}_d(t_0,u)}|,\ \forall (t_0,u) \in \mathcal{W}\}$, avec $\mathcal{W}$ l'ensemble des couples temps-nœud de départ des cascades.
— Pour chaque $|V_{\mathscr{C}_d(t_0,u)}|$ pris dans I,

— On tire un nombre $|V_{\mathscr{C}_d(t_0,u)}|$ de temps d'interaction t dans l'intervalle $[t_0, t_0 + d]$, que l'on trie ensuite dans l'ordre croissant.
— On pose $i = 0$. Puis, à chaque instant de temps t, on incrémente i. On obtient alors i en fonction de t, dont on calcule l'ASCI.

En traçant l'ASCI en fonction de la taille des cascades (non représenté) , on observe une relation linéaire entre ces deux variables. Ce constat est incompatible avec les résultats de la figure 5.2 : sur cette figure, on observe une relation parabolique entre les tailles et les ASCI des cascades. Il existe donc une dynamique d'interaction, qui ne peut être expliquée simplement par un modèle où les interactions se produisent à des instants aléatoires.

5.3.2 Modèle à deux phases

Nous avons précédemment mis en évidence l'existence d'une dynamique de diffusion commune entre les propagations de toutes tailles et de toutes ASCI, caractérisée par deux régimes distincts. Nous décidons donc de tester un modèle à deux phases : l'une où la vitesse de propagation est négligeable, et l'autre où le nombre d'infectés croît linéairement. Nous ferons référence à la première phase sous le terme *phase d'attente*, et à la seconde sous le terme phase de *croissance linéaire*.

• **Paramètres du modèle**
Nous décidons de modéliser la phase d'attente par une vitesse de propagation nulle. Le nombre d'infectés k_1 est donc constant. La deuxième phase, à croissance linéaire, débute après un certain temps, que nous appelons *temps d'attente* w. Ce temps peut être nul, autrement dit, la première phase peut être inexistante. La vitesse de propagation de cette phase est notée a. Mathématiquement, le modèle s'exprime donc de la façon suivante :

$$\begin{cases} \text{y} = k_1, & \text{si t} \leq \text{w} \\ \text{y} = a\ \text{t} + k_2, & \text{si t} > \text{w} \end{cases} \tag{5.3}$$

La dynamique de propagation d'une cascade peut être résumée aux paramètres k_1, k_2, a, et w. La figure 5.6 montre un exemple d'approximation d'une diffusion par le modèle à deux phases, lorsque sont fixés k_1 et k_2 à 1. Pour chaque cascade obtenue au chapitre précédent, nous utilisons une bibliothèque python [1] pour trouver les valeurs de a et w permettant d'obtenir la meilleure modélisation de la cascade. Pour ce faire, cette bibliothèque utilise l'algorithme de Levenberg-Marquardt. Nous calculons ensuite l'ASCI à partir de ces paramètres.

La figure 5.7 montre l'ASCI en fonction du nombre d'infectés pour les cascades mesurées et pour les résultats du modèle à deux phases. On observe que le modèle permet de trouver

1. bibliothèque ***lmfit*** : https ://pypi.org/project/lmfit/

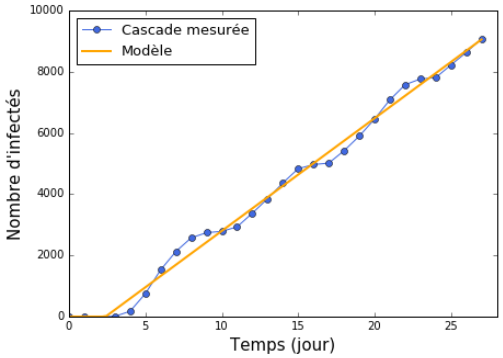


FIGURE 5.6 – Exemple de propagation d'un mois, et de l'approximation correspondante, obtenue avec le modèle à deux phases.

une allure en aile similaire aux résultats des mesures. Autrement dit, les paramètres du modèle à deux phases permettent de retrouver la diversité des ASCI correspondant à un même nombre final d'infectés.

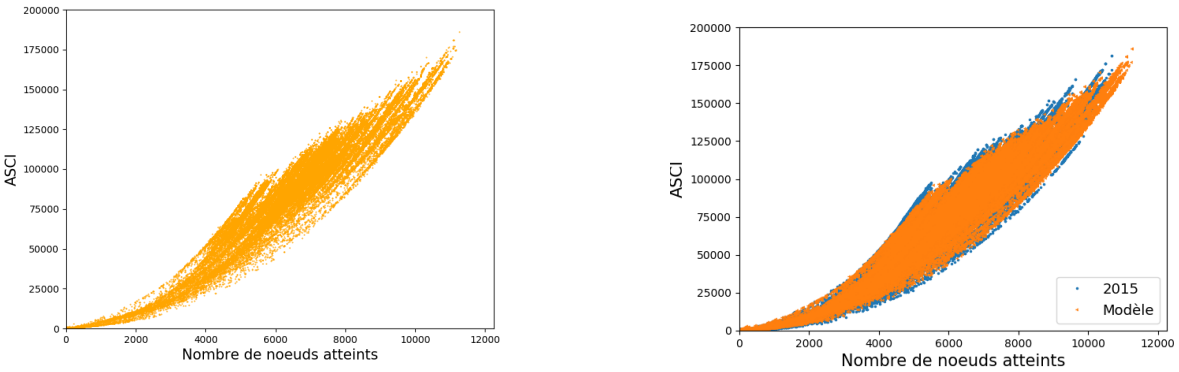


FIGURE 5.7 – ASCI en fonction des tailles des cascades de durée d'un mois, pour le modèle à deux phases (à gauche), et pour le modèle et les mesures sur les données de 2015 (à droite).

Dans notre cas, modéliser le phénomène de saturation, couramment observé dans le contexte épidémiologique, ne semble pas essentiel pour avoir une bonne représentation de la dynamique de diffusion dans le réseau. Ceci s'explique par la durée de propagation choisie (1 mois à 1 an) : celle-ci n'est pas assez longue pour qu'un phénomène de saturation apparaisse. Au contraire, sur d'autres types d'interactions, par exemple des réseaux d'appels téléphoniques [Miritello et al., 2011, Peruani and Tabourier, 2011] ou de transmission de maladies humaines [Rocha and Blondel, 2012], on observe une phase de croissance explosive (superlinéaire), puis une saturation.

Pour étudier la dynamique de propagation dans la BDNI, pour des durées de diffusion d'un mois à un an, une bonne approximation est obtenue en connaissant :
- son temps d'attente avant l'accélération de la diffusion ;
- sa pente, i.e. le coefficient directeur de la droite du nombre d'infectés en fonction du temps, représentant sa vitesse de diffusion.

5.4 Étude des caractéristiques du modèle à deux phases

Le modèle à deux phases nous permet d'obtenir pour chaque cascade son temps d'attente et sa pente lors de la phase de croissance linéaire. Nous étudions la distribution de ces variables et cherchons comment celles-ci conditionnent l'allure de la figure 5.2

5.4.1 Distribution des temps d'attente et de la pente

La figure 5.8 montre la distribution des temps d'attente pour des propagations d'un mois. La distribution obtenue est hétérogène :
- un grand nombre de propagations obtiennent un très court temps d'attente. Par exemple, 40% des cascades ont un temps d'attente strictement inférieur à 2 jours. La première phase de propagation est alors inexistante ou presque, et la diffusion débute rapidement ;
- un nombre plus restreint de propagation obtient de fortes valeurs de temps d'attente. Par exemple, $2,5\%$ des cascades ont un temps d'attente de plus de 20 jours.

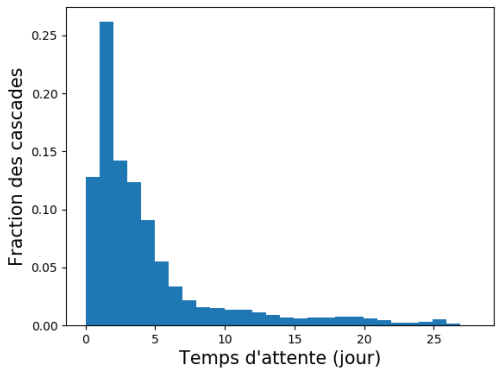


FIGURE 5.8 – Distribution des temps d'attente, obtenus sur les diffusions de 2015 de durée d'un mois.

La figure 5.9 représente la distribution des pentes pour des propagations d'un mois. On remarque que les pentes ont une distribution relativement homogène. Si l'on ne considère

pas les pentes de valeurs inférieures à 1, qui correspondent à des cascades n'atteignant pas le régime linéaire car leur taux de croissance reste équivalent à celui du temps d'attente, la pente moyenne est d'environ 252 nœuds par jour, et l'écart-type des valeurs vaut environ 53.

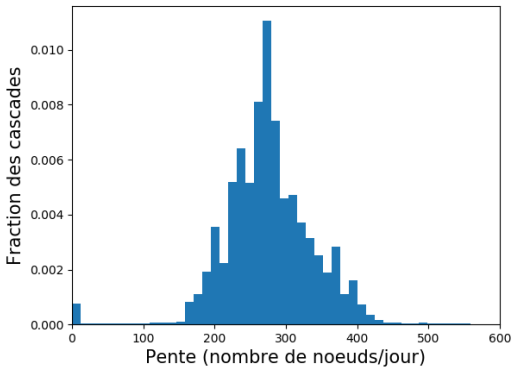


FIGURE 5.9 – Distribution des pentes des cascades d'un mois de 2015. Les valeurs sont arrondies à l'unité.

Afin de tester de quelle manière les tailles des cascades sont dépendantes des pentes, on représente en figure 5.10 l'ASCI de cascades obtenues par le modèle à deux phases mais où la pente a une valeur fixée. Nous représentons trois cas : $a \simeq \overline{a}$ (pente moyenne), $a \simeq 1,25 \cdot \overline{a}$, et $0,75 \cdot \overline{a}$. On obtient trois paraboles. C'est pourquoi, on suppose que les paraboles observées en figure 5.2 (aile réelle) sont dues à des valeurs de pente similaires entre les propagations. A partir des définitions 5.1 et 5.3 de l'ASCI et de la définition du modèle à deux phases, on exprime l'ASCI en fonction du nombre de nœuds atteints. En utilisant la relation entre le nombre de nœuds et la pente, on trouve que l'ASCI suit une équation du type :

$$\forall (t,u) \in \mathcal{W},\ ASCI_d(t,u) = \frac{1}{2a} \times (|V_{\mathscr{C}_d(t,u)}| - 1)^2 + 2d \tag{5.4}$$

5.4.2 Indépendance du temps d'attente et de la pente

Dans cette partie, on teste si les valeurs des temps d'attente et des pentes sont inter-dépendantes. Pour ce faire, on tire aléatoirement un temps d'attente pour une valeur de pente, et on calcule l'ASCI avec ces deux valeurs :

— Soit Q_a l'ensemble des valeurs de pente de cascades mesurées. Soit Q_w l'ensemble des valeurs de temps d'attente de ces mêmes cascades.

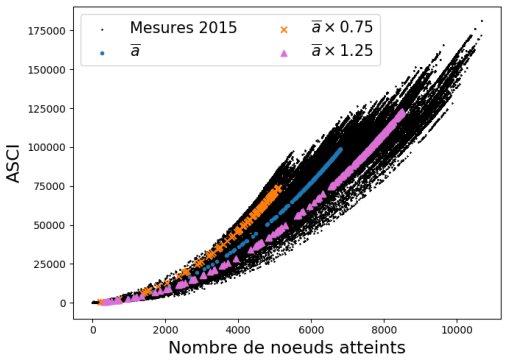


FIGURE 5.10 – Virulence en fonction du nombre d'infectés, avec 3 valeurs de pente fixées : $0,75 \cdot \overline{a}$, $\overline{a}$, et $1,25 \cdot \overline{a}$.

— Pour tout élément a de Q_a, on tire de manière aléatoire un temps d'attente w dans Q_w.
— On calcule l'ASCI des cascades générées par le modèle à deux phases avec les couples de valeurs a - w, ayant ainsi été appariées de manière aléatoire.

On constate une relation linéaire entre ces deux grandeurs. Un modèle dans lequel le temps d'attente et la pente sont tirés indépendamment ne produit pas un nuage de point en forme d'aile. Les différents paramètres du modèle ne sont pas indépendants.

La relation parabolique entre l'ASCI et le nombre de nœuds atteints s'explique par des valeurs de pente distribuées de manière homogènes. La distribution des temps d'attente est quant à elle hétérogène, et participe à la diversité de taille de propagations en fonction de la vitesse.

5.5 Impact du type des nœuds

Dans le chapitre 2, nous avons vu que les marchés, les centres de rassemblement, et certaines exploitations sont les hubs du réseau. Ces hubs pourraient être à l'origine de diffusions de plus grandes tailles, ou de plus grandes ASCI. Afin d'étudier cette hypothèse, nous représentons donc en figure 5.11 les propagations en fonction du type du nœud de départ de la diffusion, pour l'exemple de l'année 2015.

Les diffusions au départ des marchés obtiennent des ASCI localisées dans la moitié supérieure du nuage de points, ce qui signifie qu'ils sont à l'origine de propagations diffusant rapidement. De plus, on remarque qu'ils sont à l'origine de cascades de tailles supérieures à 4000 nœuds. Ainsi, les marchés sont non seulement sources de diffusions rapides, mais également de diffusions de grandes tailles. Les diffusions au départ des centres de rassem-

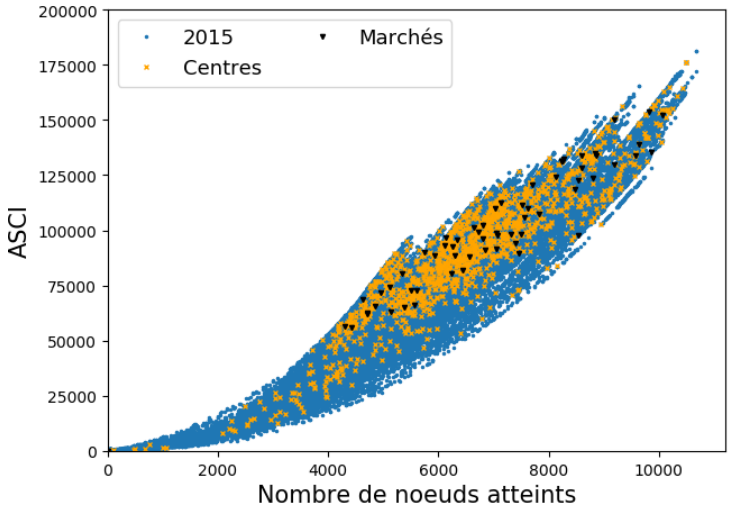


FIGURE 5.11 – ASCI en fonction du nombre d'infectés, selon le type du nœud source de la propagation (année 2015 et durée de diffusion d'un mois).

blement se concentrent également dans les valeurs hautes d'ASCI, quoique cette tendance soit moins marquée que pour les marchés. Les centres de rassemblement seraient donc également propices à la propagation rapide. Tous les types de propagations sont représentés parmi les cascades ayant pour source les élevages. Ce résultat n'est pas surprenant, étant donné que les élevages représentent l'écrasante majorité des nœuds du réseau. En revanche, lorsque l'on étudie les distributions des temps d'attente en fonction du type des nœuds, on n'observe pas de différence significative. Lorsque l'on étudie les distributions des pentes en fonction du type des nœuds, les élevages sont à l'origine d'une plus grande proportion de diffusions de faibles valeurs de pente (figure 5.12). Les pentes seraient donc un facteur expliquant les résultats plus élevés d'ASCI des cascades au départ des centres de rassemblement et des marchés.

5.6 Impact des mesures de lutte sur la dynamique de propagation

Dans le chapitre 4, nous supprimions des nœuds, des liens statiques, et des liens temporels dans le réseau. Dans cette partie, nous étudions l'impact des différentes stratégies de suppression sur les caractéristiques des propagations, analysée au travers du modèle à deux phases : leur temps d'attente avant l'accélération de la diffusion, et leur pente (la vitesse du régime de croissance linéaire).

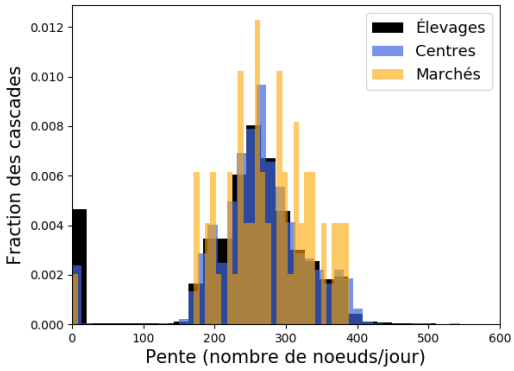


FIGURE 5.12 – Distribution des valeurs de pentes de 2015 en fonction du type de l'exploitation source de la diffusion, de durée d'un mois.

5.6.1 Impact sur la relation entre le nombre d'infectés et l'ASCI

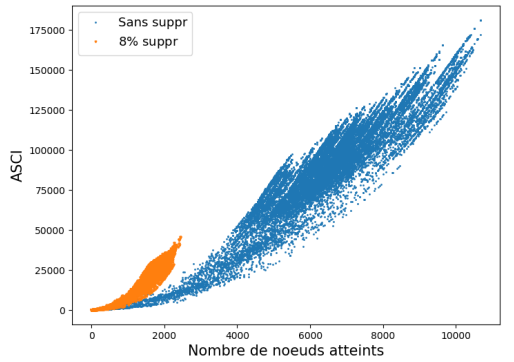
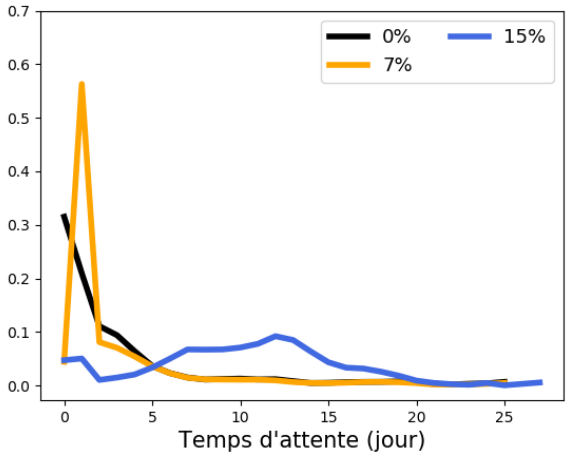


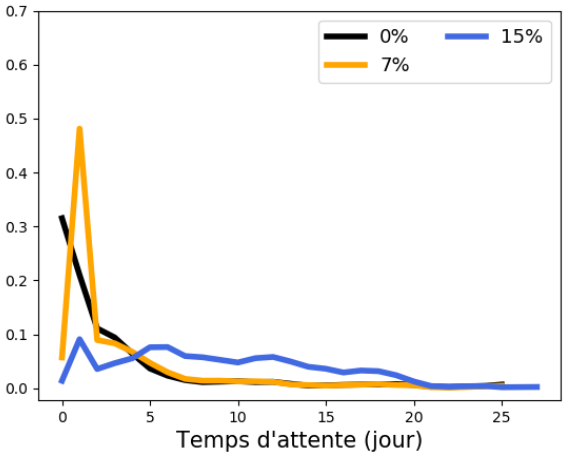
FIGURE 5.13 – Impact sur l'ASCI de la suppression de 8% des liens temporels selon leur nombre d'occurrences dans les cascades de 2015 de durée d'un mois, en fonction du nombre de nœuds atteints par les diffusions.

La figure 5.13 montre un exemple d'impact d'une stratégie de suppression sur l'ASCI en fonction des tailles des cascades, ici avec la méthode $N_{occ}(t, u, v)$ (définition 4.1), comptant les occurrences des liens temporels dans les cascades. On observe que les nombres de nœuds atteints et les ASCI correspondantes diminuent, en conservant leur relation parabolique et une enveloppe en forme d'aile, observée dans la partie précédente. Nous pouvons donc utiliser le même modèle à deux phases pour étudier la vitesse et le temps d'attente avant accélération des propagations, dans le cas où des stratégies de suppression sont mises en œuvre.

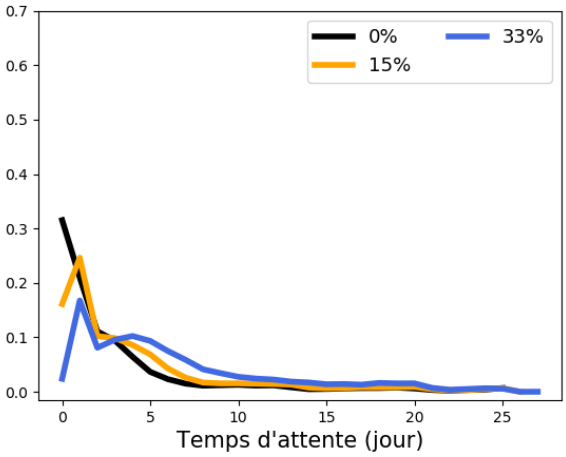
5.6.2 Impact des suppressions sur le temps d'attente



(a) Suppression selon les $N_{occ}(t, u, v)$



(b) Suppression selon les $N_{occ}(u, v)$



(c) Suppression selon les k_{out}

FIGURE 5.14 – Distribution des temps d'attente d'un échantillon de diffusions obtenues avec le modèle à deux phases en fonction du nombre de liens temporels supprimés, selon les stratégies : nombre d'occurrences des liens temporels dans les cascades $N_{occ}(t, u, v)$ (a), nombre d'occurrences des liens statiques $N_{occ}(u, v)$ (b), et degré sortant des nœuds k_{out} (c).

La figure 5.14 présente les distributions des temps d'attente selon la stratégie de ciblage considérée, pour des suppressions exprimées en pourcentages de liens temporels (définition 4.2). Les suppressions de liens statiques et temporels selon leurs nombres d'occurrences dans les cascades (définition 4.1) ont un impact similaire sur les distributions :

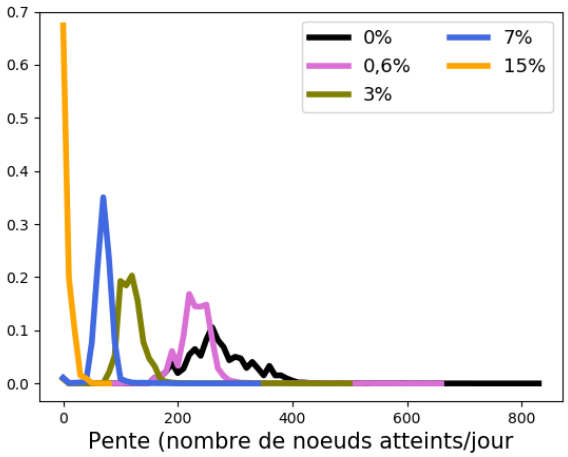
jusqu'à 7% de liens supprimés, les temps d'attente de 1 jour augmentent subitement en nombre, pour finalement obtenir, avec 15% de suppressions, une distribution où les temps d'attente courts sont en plus faible proportion par rapport aux temps d'attentes d'une dizaine de jours. On passe donc de distributions très hétérogènes à des distributions qui tendent à devenir plus uniformes. Au contraire, à 15% de suppressions, les temps d'attente courts sont toujours présents de manière majoritaire lorsque l'on supprime les nœuds selon leur degré sortant. Or, des temps d'attente allongés sont signes de propagations mettant plus de temps à atteindre la phase de croissance linéaire du nombre d'infectés en fonction du temps. Les suppressions de liens statiques et temporels selon leurs nombres d'occurrences dans les cascades entraîneraient donc un ralentissement plus fort des diffusions, par rapport à la suppression selon le degré sortant des nœuds.

5.6.3 Impact des suppressions sur la pente

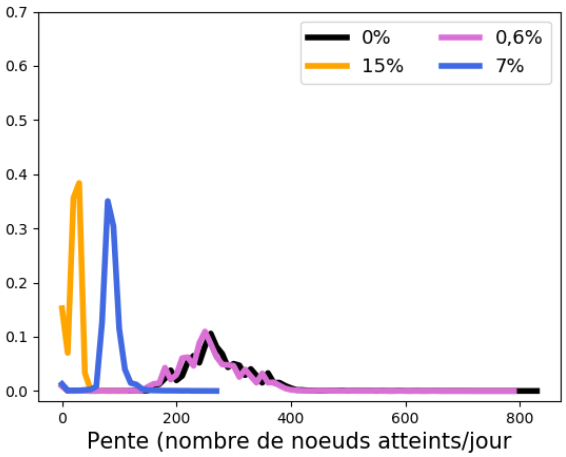
La figure 5.15 présente les distributions des pentes selon la stratégie de ciblage considérée, et le nombre de suppressions effectuées (exprimé en pourcentage de liens temporels). Pour le degré sortant, on observe qu'augmenter le nombre de suppressions décale la distribution des pentes vers de plus petites valeurs. Les pics semblent peu varier d'amplitude. Concernant la suppression de liens statiques ou temporels selon leurs nombres d'occurrences dans les cascades, on observe non seulement un décalage des distributions vers de plus faibles valeurs de pente, mais également l'augmentation de la taille des pics.

Ainsi, supprimer des nœuds selon leur degré diminue les pentes des propagations, dans des proportions similaires pour chacune d'elles. Les propagations sont donc globalement plus lentes, mais ne semblent pas être interrompues par cette mesure, car on n'observe pas d'augmentation nette de la proportion des pentes de faibles valeurs par rapport aux autres valeurs. Au contraire, la suppression de liens permet à la fois de ralentir les propagations, en diminuant leurs pentes, mais semble également les stopper. Les stratégies de suppression de liens et de nœuds ont donc un impact qualitativement différent sur les diffusions, qui peut expliquer la plus grande efficacité des méthodes de suppression de liens, observée au chapitre 4.

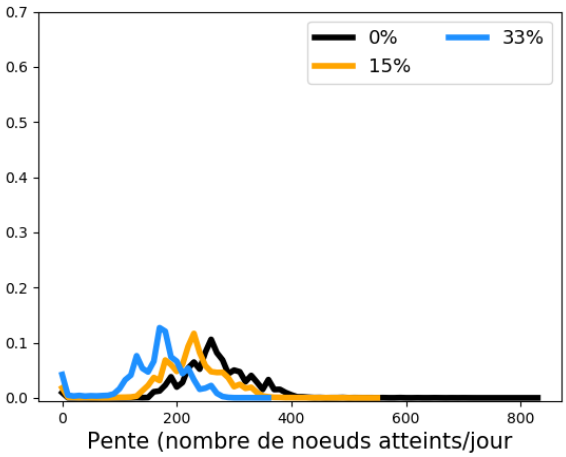
L'étude de l'ASCI en fonction du nombre de nœuds atteints montre que les suppressions n'ont pas d'impact sur la relation entre ces deux variables. La modélisation des propagations avec le modèle à deux phases montre que les suppressions d'éléments importants dans les diffusions permettent de réduire leur pente (vitesse), et de rallonger le temps d'attente avant l'accélération de la diffusion.



(a) Suppression de liens temporels



(b) Suppression de liens statiques



(c) Suppression de noeuds (degré sortant)

FIGURE 5.15 – Distribution des pentes d'un échantillon de diffusions obtenues avec le modèle à deux phases, selon le mode de suppression : selon l'occurrence des liens temporels dans les cascades (a), selon l'occurrence des liens statiques (b), et le degré sortant des nœuds (c).

Conclusion

Les cascades présentent une diversité de tailles et d'ASCI importantes. L'étude de leurs profils de propagation permet de montrer qu'elles répondent néanmoins à la même dynamique, approximativement à deux phases linéaires. Ainsi, elles ne présentent pas de phénomène de saturation, contrairement aux profils classiquement observés dans la litté-

rature. Les durées d'observation des diffusions sur la BDNI, classiquement utilisées dans la littérature (de un mois à un an), sont plus courtes que la période nécessaire pour observer un phénomène de saturation. Ce n'est pas le cas d'autres types de données de la littérature, par exemple les appels téléphoniques : les durées d'observations et d'apparition d'un phénomène de saturation sont comparables. On observe ainsi toutes les étapes de la diffusion.

Les cascades, décrites par le modèle à deux phases, sont caractérisées par deux paramètres : une pente et un temps d'attente. Lorsqu'elles partagent des valeurs de pente proches, les cascades ont des tailles et des ASCI reliées par une relation parabolique. Les pentes sont donc un facteur important expliquant l'existence d'une diversité de valeurs d'ASCI pour une même taille de cascade.

Les différences d'efficacité des stratégies d'identification de nœuds et de liens importants, pour réduire la taille des cascades, s'expliquent par une action différente sur les pentes et les temps d'attentes. Notamment, la suppression de liens statiques ou temporels selon leurs scores N_{occ} permet de stopper les propagations : les cascades concernées présentent alors des valeurs de pente nulles. Les autres cascades sont ralenties, c'est-à-dire que leurs valeurs de pente diminuent. Au contraire, la suppression de nœuds aboutit à la diminution des valeurs de pente, dans des proportions équivalentes pour chaque cascade, sans toutefois parvenir à stopper les diffusions.

Chapitre 6

Conclusion et Perspectives

Sommaire

Dans ce chapitre, nous présentons les conclusions tirées de nos travaux de thèse. Puis nous détaillerons certaines perspectives dégagées, qui nous semblent parmi les plus importantes.

6.1 Conclusion

Le but de cette thèse était d'utiliser l'information temporelle pour :
— évaluer l'impact des propagations sur le réseau d'échanges de bovins en France ;
— améliorer la détection d'exploitations et d'interactions à risque pour le réseau.

Pour répondre à ces questions, nous avons tout d'abord défini dans le chapitre 3 un protocole de comparaison des mesures des tailles de propagations dans les graphes et les

réseaux temporels. Nous avons alors montré que prendre en compte la temporalité des interactions révèle l'existence d'une diversité de tailles de propagations. Au contraire, les propagations de tailles intermédiaires ne sont pas détectées et voient leur taille surestimée lorsque l'ordre chronologique des interactions est négligé.

De plus, l'accès à l'information temporelle sur les échanges nous permet de mesurer non seulement la taille des propagations, mais également leur vitesse. Nous proposons dans le chapitre 5 d'exprimer cette vitesse sous la forme de l'ASCI (l'aire sous la courbe du nombre d'infectés en fonction du temps).

Dans le cas de la BDNI, la mesure de l'ASCI montre que peu de scénarios de cascades existent. L'étude de ces scénarios de diffusion nous amène à proposer un modèle d'accessibilité des nœuds, caractérisé par deux phases :

— une première phase d'attente, de vitesse négligeable, durant un certain temps, nommé le temps d'attente ;
— et une phase de croissance linéaire, caractérisée par sa pente.

Compte-tenu de l'existence d'un unique scénario de propagation, la mesure des pentes permet une caractérisation adéquate de la vitesse des diffusions. Cependant, sur d'autres données ou avec un protocole différent, des scénarios différents pourraient exister, et l'ASCI permettrait de les distinguer.

Par ailleurs, nous avons testé dans le chapitre 4 l'efficacité de différentes stratégies pour identifier les nœuds et liens importants pour la propagation. Nous montrons que supprimer un pourcentage de nœuds jugé comme faible peut correspondre à la suppression d'un nombre très important de liens. Exprimer les suppressions en pourcentages de nœuds donne donc l'impression qu'une stratégie est mois coûteuse pour le réseau qu'elle ne l'est réellement, contrairement à l'utilisation du volume de suppression (nombre de liens temporels supprimés). De plus, l'utilisation du volume de suppression permet de facilement comparer les stratégies de suppression de nœuds et de liens statiques et temporels.

En outre, nous proposons dans le chapitre 4 une stratégie intégrant l'information temporelle pour identifier les interactions jouant un rôle majeur dans la diffusion sur le réseau. Elle consiste à compter le nombre de liens statiques apparaissant dans les cascades. Elle permet une réduction des tailles des cascades à faible coût (*i.e.* faible volume de suppressions). De plus, les liens qu'elle identifie comme importants ont tendance à le rester au cours du temps. Autrement dit, le classement des liens importants réalisé à une période donnée, permettra une identification des éléments importants sur une période ultérieure qui sera plus efficace que les stratégies de la littérature testées. Ainsi, les liens importants peuvent être identifiés lorsqu'il existe un délai de transmission de l'information, entre les dates où les échanges d'animaux ont lieu et leur enregistrement dans la BDNI.

Ces travaux ouvrent de nombreuses perspectives. Tout d'abord, la dynamique de propagation observée, à deux phases linéaires, ne correspond pas à ce qu'on aurait pu attendre : il est usuel de considérer que les diffusions finissent par saturer. Si ce modèle est une bonne approximation de la dynamique sur des échelles de temps d'une année ou moins, on peut donc s'interroger sur son amélioration pour l'adapter à des durées de mesures qui couvrent l'apparition du phénomène de saturation.

Ensuite, on souhaite étudier plus précisément la structure des cascades : partagent-elles des chemins de propagation et à quel point ? Quel rôle jouent les marchés et centres de regroupement ? En particulier, quels types de nœuds et de liens permettent l'accélération de la propagation ?

Enfin, on souhaite étudier le rôle des propriétés structurelles et temporelles du réseau temporel dans la diffusion. Pour ce faire, ces caractéristiques doivent être éliminées spécifiquement pour étudier les potentielles différences dans la dynamique de propagation. Cette élimination se fait au moyen de la *randomisation* (rendre aléatoire) sélective des propriétés du réseau temporel.

Nous détaillons dans les prochaines parties ces perspectives clés et les travaux exploratoires que nous avons menés les concernant.

6.2 Profils de propagation

Nous revenons ici sur les perspectives dégagées de l'étude des profils de propagation et de leur modélisation (chapitre 5). Notamment, nous cherchons des pistes d'études afin d'expliquer l'absence de saturation dans des propagations de moins d'un an, alors que cette caractéristique est souvent décrite dans le contexte de l'épidémiologie [Barthélemy et al., 2004].

6.2.1 Diffusions sur longues périodes

Au chapitre 5, notre étude des profils de propagation porte sur des diffusions de durée inférieure à un an. Nous avions observé qu'un modèle à deux phases permet de reconstituer les principales caractéristiques de la diffusion, et constitue donc une bonne approximation à cette échelle d'observation.

La figure 6.1 montre des propagations de durée de 5 ans de grandes tailles. Contrairement aux propagations d'un an ou moins, on observe ici que la valeur de la pente diminue au cours de la diffusion, ce qui traduit l'apparition d'un phénomène de saturation. De

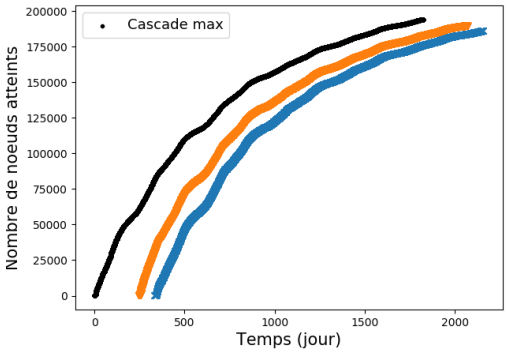


FIGURE 6.1 – Nombre d'infectés en fonction du temps pour la propagation de durée de 5 ans de taille maximale et deux autres cascades de grandes tailles.

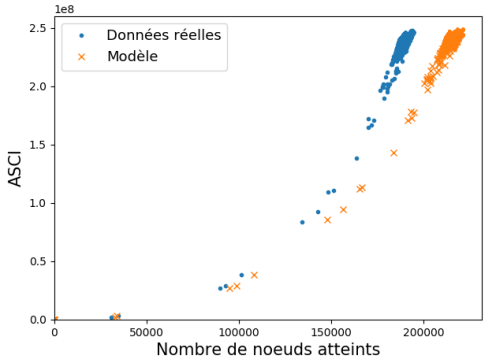


FIGURE 6.2 – ASCI en fonction des tailles de cascades, pour des propagations de 5 ans (points bleus) et pour les résultats correspondants obtenus avec le modèle à deux phases (croix oranges).

plus, on n'observe pas ici de première phase de propagation plus lente, approximativement constante. La modélisation des diffusions par un modèle à deux phases est donc remis en question pour des propagations de plus longue durée qu'un an. Ainsi, la figure 6.2 montre que le modèle à deux phases conduit à une modélisation de la diffusion inadaptée pour cet échantillon de propagations de 5 ans, de par de forts écarts entre les valeurs des couples taille cascade - ASCI. Le modèle évalue systématiquement des valeurs d'ASCI inférieures à celles des cascades mesurées de taille supérieure à 100 000 nœuds. À taille fixée, les vitesses des cascades sont donc sous-estimées par le modèle, en raison de la non prise en compte de l'apparition du phénomène de saturation.

Deux pistes d'études se dégagent du constat de l'apparition d'un phénomène de saturation pour des diffusions de plus longue durée :

— parvenir à caractériser la bascule entre ces deux comportements de diffusion (linéaire

et saturant) ;
— prendre en compte la saturation dans la modélisation des propagations, en complexifiant les équations du modèle à deux phases.

Par ailleurs, on remarque que la diversité des valeurs de couples taille cascade - ASCI est plus faible pour cet échantillon de cascades de 5 ans (figure 6.2), que pour les cascades d'un mois (figure 5.2). Non seulement les valeurs d'ASCI sont moins diverses pour une taille de cascade donnée, mais les points des mesures ont tendance à se concentrer vers les valeurs extrêmes des tailles de cascades. On peut supposer que l'augmentation de la période de diffusion permet aux cascades de se rejoindre, et de partager des chemins de propagation similaires. Les tailles de cascades et les valeurs d'ASCI obtenues sont alors plus proches que dans le cas de propagation de faible durée. Nous reviendrons sur cette hypothèse en partie 6.3.

6.2.2 Absence de saturation et renouvellement des nœuds

Nous avons étudié en partie 2.3.3 la dynamique d'activation des nœuds et des liens sur des flots de liens de durée d'un an. Nous avons notamment observé que l'ensemble des nœuds s'activant pour un achat croît de manière approximativement linéaire au cours du temps (figure 2.4). Durant une période d'un an, on rencontre donc régulièrement des nœuds qui sont atteints pour la première fois.

Or, nous avons observé au chapitre 5 que les propagations peuvent être approximées par un modèle à deux phases. On peut se demander si la croissance linéaire du nombre de nœuds dans la cascade n'est pas une conséquence de l'activation de nouveaux nœuds au cours du temps. Pour confirmer ou infirmer cette hypothèse, nous traçons des profils de propagation, sans prendre en compte l'activation de nouveaux nœuds au cours du temps.

Soient $L = (T, V, E)$, $T = [t_\alpha, t_\omega]$ et un temps $t_f \in T$. $v \in V$ est dit nouveau si $\nexists t \in [t_\alpha, t_f]$ tq (t, v, u) ou $(t, u, v) \in E$. $v \in V$ est dit ancien si $\exists t \in [t_\alpha, t_f]$ tq (t, v, u) ou $(t, u, v) \in E$. Une propagation est dire restreinte si la cascade correspondante ne comprend que des nœuds anciens sur la période $[t_\alpha, t_f]$.

Le but est de prendre en compte, dans la représentation des profils de propagation, uniquement les nœuds ayant été actifs (*i.e.* ayant été impliqué dans un échange) avant une date fixée, c'est-à-dire les nœuds anciens de la période définie. Ainsi, l'activation de nouveaux nœuds n'est plus prise en compte après cette date. Pour ce faire, nous tirons aléatoirement des cascades. Nous traçons au cours du temps les profils des propagations restreintes aux nœuds anciens sur la période définie.

Par exemple, la figure 6.3 montre une propagation de 5 ans, où les nœuds infectés ne

sont pris en compte dans le tracé de la courbe que s'ils ont été actifs les premiers 2,5 ans des données (à gauche, courbe bleue). Nous n'observons pas de différence qualitative entre cette courbe et le nombre d'infectés au cours du temps sans modification. Même en diminuant la durée de la période de définition des nœuds anciens (par exemple 6 mois, à droite de la figure), la différence reste minime : dans cet exemple, plus de 80% des nœuds atteints ont été actifs durant les premiers 6 mois de la diffusion de 5 ans. De même, la figure 6.4 montre que près de 60% des nœuds atteints par une diffusion d'un an au départ du 1er janvier ont été actifs les 50 premiers jours de l'année 2015. Dans ces exemples, les nœuds infectés ont donc été actifs avant leur date d'infection. Les nœuds atteints par ces cascades sont donc majoritairement des nœuds anciens. La différence entre le profil linéaire de la propagation d'un an, et le profil non linéaire de la propagation de 5 ans, n'a donc pas pour principale explication l'activation de nouveaux nœuds au cours du temps selon une dynamique linéaire.

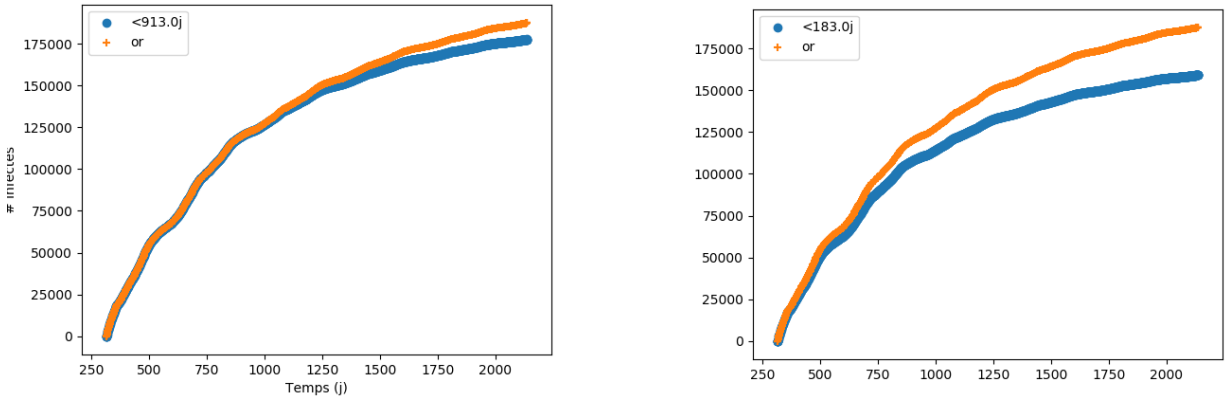


FIGURE 6.3 – Nombre d'infectés au cours du temps pour une propagation de 5 ans (courbes oranges), et pour cette même propagation où seuls les nœuds actifs les premiers 2,5 ans (à gauche) ou les premiers 6 mois (à droite) sont tracés.

6.3 Superposition des cascades

La compréhension du lien entre la mesure des tailles des cascades et de l'ASCI a été le sujet du chapitre 5. Nous avons pu déterminer l'équation reliant ces deux grandeurs (équation 5.4), selon le modèle de représentation des cascades.

Toutefois, la première hypothèse que nous avions eue à la vue de la forme parabolique de l'ASCI en fonction du nombre d'infectés (figure 5.2) était que les paraboles étaient dues à des propagations empruntant des chemins communs. En effet, des cascades débutant à des instants de temps et des nœuds sources différents peuvent partager des chemins tout

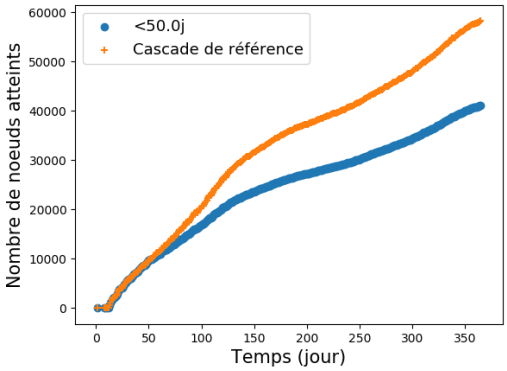


FIGURE 6.4 – Nombre d'infectés au cours du temps pour une propagation d'un an choisie aléatoirement (courbe orange), et pour cette même propagation où seuls les nœuds actifs les premiers 50 jours sont tracés.

en atteignant un nombre de nœuds différents, du fait de la borne temporelle du calcul des cascades, et également du fait que les chemins ne sont pas identiques selon la source. Il pourrait exister dans le réseau un ensemble de chemins empruntés par de nombreuses cascades, formant une sorte de colonne vertébrale du réseau. On peut alors supposer qu'il existe un lien entre les tailles et les ASCI des cascades partageant des chemins. Plus les cascades empruntent des chemins similaires, plus ce lien serait fort.

Selon la définition présentée en partie 1.2.3.1, une cascade $\mathscr{C}_d$ de durée d est un ensemble de liens temporels. Pour mesurer la ressemblance entre deux cascades, nous proposons l'utilisation de l'indice de Jaccard (équation 4.3) sur ces deux ensembles de liens.

Un protocole simple de comparaison des cascades consiste à calculer l'indice de Jaccard sur tous les couples de cascades, pour évaluer à quel point elles partagent une partie commune. Dans cette thèse, chaque nœud du flot de liens ($L = (T, V, E)$) a été source d'une propagation. Comparer les cascades deux à deux entraîne alors une complexité de calcul élevée. Par ailleurs, une fois les indices de Jaccard obtenus, une manière de procéder consiste à fixer un seuil, à partir duquel les propagations seront jugées comme similaires. Ce seuil devra être suffisamment élevé pour effectivement représenter la similarité entre les cascades, mais cependant être suffisamment bas pour qu'un nombre suffisant de cascades puissent être comparées. Une fois les indices de Jaccard calculés, on vérifie si les cascades jugées similaires correspondent effectivement à la même parabole.

Nous réalisons une version très préliminaire de ce protocole, en évaluant la similarité d'une cascade de 2015 d'une durée d'un mois, choisie aléatoirement, avec toutes les autres cascades de 2015. Nous représentons en figure 6.5 toutes les cascades qui obtiennent un indice de Jaccard supérieur à $0,6$. On remarque que les couples taille - ASCI des cascades

sont approximativement alignés, à l'exception de quelques cascades. Il pourrait donc exister une certaine similarité, et donc un partage de chemins entre cascades. Ces mesures préliminaires sont insuffisantes pour pouvoir conclure sur cette question.

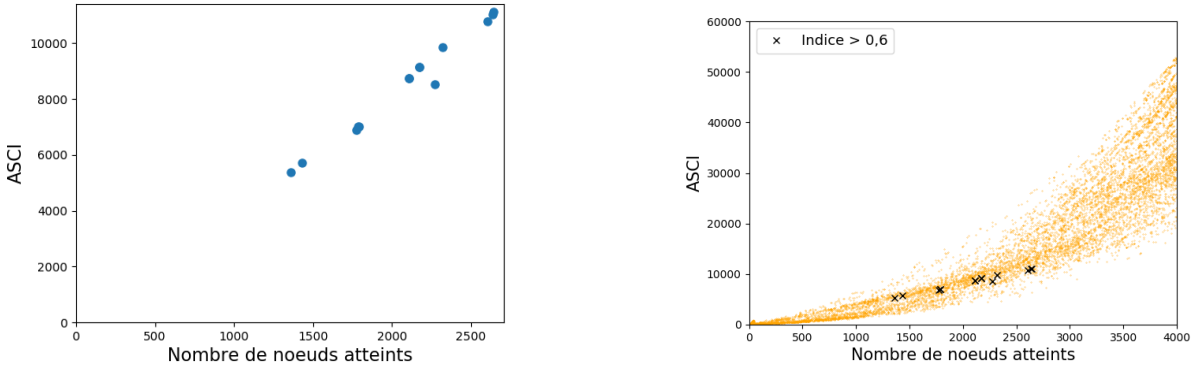


FIGURE 6.5 – ASCI en fonction du nombre de nœuds atteints, pour des cascades d'un mois de 2015, obtenant un indice de Jaccard supérieur à $0,6$ lorsque comparé à une cascade choisie aléatoirement et atteignant 1788 nœuds. A droite, les points sont replacés dans l'ensemble des cascades de taille inférieure à 4000 nœuds.

L'étude des chemins communs et la recherche de la colonne vertébrale du réseau est donc une perspective intéressante à cette thèse. Le protocole proposé devrait être exploré de manière plus exhaustive en examinant un plus grand nombre de cascades, et en utilisant d'autres valeurs de seuil. De plus, il faudrait étudier les pentes des cascades identifiées comme similaires, afin de voir s'il existe un lien entre ces deux caractéristiques.

6.4 Stratégies d'identification

Nous avons vu au chapitre 4 une grande variété de stratégies d'identification des nœuds et des liens clés dans les diffusions sur le réseau. Nous revenons tout d'abord sur la place des marchés et des centres de rassemblement dans les classements d'importance. Nous proposons également d'autres pistes d'amélioration, en lien avec nos mesures des occurrences dans les cascades (définitions 4.1). Puis, nous revenons sur la question, abordée en fin de chapitre 4, de la prévision des liens temporels à supprimer.

6.4.1 Importance des marchés et des centres de rassemblement

Nous avons vu dans la littérature l'importance des centres et des marchés dans la structure du réseau. Par exemple, comme détaillé au chapitre 4, [Rautureau, 2012] évalue le nombre de marchés ou de centres de rassemblement nécessaire pour faire disparaître la GSCC. De plus, nous avons vu au chapitre 2 l'importance de ces types de nœuds dans les GSCC, et dans les échanges de par leur degré et leur activité importantes.

Dans le chapitre 5, nous avons montré que deux phases se succèdent au début des diffusions : une lente, et une rapide. Il serait intéressant de mesurer la proportion de chaque type de nœuds et de liens dans ces deux phases. Ainsi, l'on pourrait déterminer si les centres et les marchés jouent également un rôle particulier dans l'apparition de la phase d'accélération de la diffusion. Plus spécifiquement, on pourrait déterminer si un type de liens est plus fréquemment impliqué dans l'apparition de la phase d'accélération, notamment ceux impliquant les marchés et les centres de rassemblement. Si les marchés et centres font effectivement partie des liens responsables de l'accélération de la propagation, il sera alors intéressant d'observer si ce sont principalement les liens sortant ou entrant.

6.4.2 Motifs dans les flots de liens

Nous avons vu au chapitre 4, que [Kao et al., 2006] proposent de supprimer les élevages qui permettent de créer un chemin entre deux marchés différents. Ces élevages permettent en effet de relier différentes communautés en une même composante fortement connexe. Les auteurs évaluent donc si leur suppression permet de réduire efficacement la taille de la GSCC. Ce faisant, les auteurs énumèrent donc le motif A→B→C (i.e. A vend un animal à B qui vend un animal à C) selon la nature des nœuds A, B, et C, en l'occurrence marchés pour A et C et élevage pour B. Une étude similaire pourrait être menée, lorsque les exploitations A et C sont des centres de rassemblement. En outre, d'autres combinaisons de types de nœuds, voire de types de liens, pourraient être étudiées.

Par ailleurs, le motif de graphe A→B→C est statique, et nous avons mis en évidence que l'ordre chronologique des interactions jouait un rôle fondamental. Rechercher des motifs temporels déclenchant l'accélération serait pertinent. [Holme and Saramäki, 2012, Kovanen et al., 2011] décrivent les motifs dans le cas des réseaux temporels. Tout d'abord, il faudrait définir le ou les motifs d'intérêt pour la diffusion dans le réseau d'élevages. Une fois définis, on pourrait évaluer l'importance d'un motif dans la diffusion, en le supprimant. Plusieurs protocoles pourraient être testés. Tout d'abord, on pourrait supprimer les motifs intégrant certains types de nœuds (notamment les marchés ou les centres de rassemblement) ou de liens. Ensuite, les motifs pourraient être recherchés dans les cascades, et leur fréquence d'apparition mesurée en fonction du temps. Les motifs ayant la fréquence la plus élevée

juste avant la phase d'accélération de la diffusion pourrait être identifiés, en vue d'être supprimés. Nous pensons en effet qu'il existe des motifs temporels dont l'existence serait corrélée avec le phénomène d'accélération de la croissance des cascades, et nous souhaitons les identifier.

6.4.3 Suppression des liens temporels selon leur distance au temps d'attente

Une piste intéressante pour caractériser les éléments responsables de l'accélération de la propagation consisterait à étudier leur distance au temps d'attente. Dans cette partie, nous nous concentrons sur le cas des liens, compte tenu des bons résultats des méthodes de suppression de ce type (chapitre 4). Nous évaluons donc, pour un lien temporel (t, u, v), sa distance $|t - w|$ au temps d'attente w, w étant obtenu grâce au modèle à deux phases (chapitre 5).

La diffusion accélérant à partir du temps d'attente, il est probable que certains liens aux caractéristiques particulières soient atteints à ce moment là, et soient responsables de ce phénomène. Par exemple, on suspecte fortement les liens (t, u, v) où u ou v sont soit un marché, soit un centre de rassemblement, de faire partie de ces liens accélérateurs de la propagation. Nous souhaitons donc mettre en place une mesure prenant en compte la distance au temps d'attente pour juger de l'importance des liens. Cependant, nous pensons que le nombre d'occurrences du lien temporel dans les cascades doit également être pris en compte dans la mesure. Il faut notamment éviter de surestimer l'importance des liens qui apparaissent dans les cascades de très petites tailles, même s'ils ont lieu à des temps proches du temps d'attente. C'est pourquoi, nous calculons :

$$D(t, u, v) = \frac{N_{occ}(t, u, v)^2}{\sum_{\mathscr{C}_d} |t - w|}$$

Un lien ayant un score D élevé est donc un lien infecté à un temps t proche du temps d'attente w. De plus, un score élevé traduit une plus grande participation dans les diffusions et donc un rôle plus important.

Les liens sont ensuite classés et supprimés selon leur score décroissant. On mesure le ratio de la taille de la plus grande cascade avec suppression, sur celle sans suppression. La figure 6.6 montre ce ration en fonction du nombre de suppression. On remarque que la suppression selon la distance au temps d'attente ne permet pas de diminuer plus efficacement les tailles des cascades que le nombre d'occurrences des liens temporels dans les cascades, méthode dont elle est issue. Si le score D proposé n'est pas suffisamment efficace pour stopper les propagations, nous pensons toutefois qu'identifier les liens temporels étant

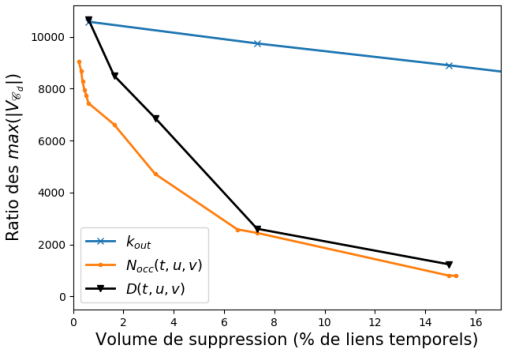


FIGURE 6.6 – Tailles de cascades maximales sur la taille maximale sans suppression, en fonction du nombre de liens temporels supprimés.

atteints à un instant proche du temps d'attente reste une question pertinente.

6.4.4 Vers la prédiction des interactions

Dans le chapitre 4, nous évoquons la difficulté de supprimer des liens temporels dans le cadre d'études prédictives : les dates des liens utilisés pour les calculs ne correspondent évidemment pas à un instant de temps des données futures. Dans cette partie, nous présentons une piste d'étude pour prédire, à partir d'informations sur les échanges passés, l'importance de liens temporels dans le futur.

Le principe de cette étude consiste à évaluer, pour chaque lien temporel rencontré au cours du temps, sa capacité à connecter des parties distinctes du réseau temporel. Pour ce faire, nous avons calculé les chaînes d'infection entrantes [1] des nœuds au fur et à mesure de leur rencontre sur les cascades :

— on rencontre un lien temporel reliant les nœuds A et B au temps t au cours de la simulation de la diffusion ;
— on calcule la chaîne d'infection entrante à partir de B au temps t ;
— de même, on calcule la chaîne d'infection entrante à partir de B au temps t, mais cette fois-ci sans prendre en compte son interaction avec A au temps t ;
— on calcule la différence entre les deux chaînes d'infection entrantes. Nous pensons en effet que ce score traduit la capacité du lien à connecter des régions différentes du réseau.

On obtient pour la période de temps étudiée des scores pour chacun des liens temporels

1. Chaîne d'infection entrante = nombre de nœuds pouvant atteindre le nœud étudié, en respectant l'ordre chronologique des interactions. On peut donc considérer que c'est la taille d'une cascade lorsque l'on remonte dans le temps.

rencontrés.

Afin de vérifier si les liens temporels de plus hauts scores jouent un rôle important pour la diffusion sur la période étudiée, on simule des propagations où les liens sont supprimés selon leur score décroissant. Les résultats préliminaires (non représentés) ne montrent pas une plus grande efficacité que la stratégie de suppression des nœuds selon leur degré ou leur centralité d'intermédiarité dans le réseau. La suppression de liens statiques selon leur occurrence dans les cascades reste nettement la méthode la plus efficace. Une explication possible est que l'asymétrie des contacts (voir partie 2.3.4) fait qu'un lien important pour les chaînes d'infection entrantes peut avoir un rôle mineur pour les chaînes sortantes.

Par ailleurs, prédire le risque que représenterait un lien temporel futur suppose de savoir prédire quels liens sont effectivement susceptibles d'apparaître et quand. La littérature regorge de méthodes de prédiction d'apparition des liens dans un réseau. Mais dans le cadre de notre étude, l'enjeu est de trouver des méthodes qui prennent en compte à la fois la structure et l'information temporelle sur les interactions. Nous souhaiterions donc réaliser une prédiction de liens temporels dans un réseau dynamique, qui est encore un domaine peu exploré. Elle peut être réalisée sur une séquence de graphes comme par exemple dans [Rahman et al., 2018]. Cependant un formalisme comme celui des flots de liens pourrait être plus adapté car il intègre les aspects structurels et temporels sans exiger le choix d'une échelle de temps d'analyse.

6.5 Impact des caractéristiques structurelles et temporelles sur la propagation

Le but de la randomisation est de comprendre l'impact des propriétés structurelles et temporelles du réseau sur la diffusion. La randomisation consiste en effet à détruire certaines de ces propriétés sélectivement pour ensuite pouvoir observer les changements ayant lieu dans la diffusion. Dans la littérature, de nombreuses études y font appel. [Miritello et al., 2011] et [Crépey and Barthélemy, 2007] utilisent par exemple des modèles de randomisation des étiquettes temporelles des liens pour étudier le rôle des corrélations temporelles entre interactions sur la diffusion. [Scholtes et al., 2014] ont recours à un modèle de randomisation pour étudier les relations de causalités entre liens successifs, liens formant le motif $A \rightarrow B \rightarrow C$, avec A, B et C des nœuds. Des revues des modèles existants sur les réseaux temporels ont été proposées par [Holme and Saramäki, 2012] et [Gauvin et al., 2018].

En ce début de travail sur la randomisation, nous rapportons les premières mesures exploratoires obtenues avec le modèle RT (pour *Random Time*) [Holme and Saramäki, 2012]. Le modèle RT consiste à remplacer tous les temps d'interaction des liens temporels

du réseau par un temps tiré aléatoirement entre le temps initial t_α et le temps final t_ω des données. Ce modèle détruit toutes les caractéristiques temporelles des interactions, notamment les corrélations entre échanges et le *daily pattern*. Ce dernier terme fait référence à un cycle régulier dans les échanges. Dans notre cas, ce cycle est hebdomadaire, comme nous l'avons vu partie 2.3.7.

6.5.1 Élimination progressive des caractéristiques du réseau

La difficulté de l'étude d'un réseau randomisé est qu'il n'est pas toujours facile d'identifier quelle caractéristique éliminée par le modèle est à l'origine de quelles répercutions sur la dynamique de propagation. Le modèle RT n'y fait pas exception, comme nous l'avons vu précédemment.

Une idée pour mieux identifier les caractéristiques affectées par le modèle est d'avoir recours à une randomisation progressive du réseau. Autrement dit, en éliminant progressivement certaines caractéristiques du réseau, de manière à tendre vers le modèle, on espère détecter, de manière qualitative, à quel moment les changements s'opèrent. Par exemple, [Karsai et al., 2011] montrent que détruire les corrélations temporelles entraîne une accélération de la propagation suivant un modèle SI. On pourrait se demander à partir de quel niveau de randomisation cette accélération est observable.

A partir du modèle RT, nous envisageons deux manières différentes de randomiser progressivement le réseau, que nous nommons $RT\%$ et $RT\Delta$.

- **RT**$\%$

Avec cette approche, on passe progressivement du réseau réel à un réseau aléatoire en randomisant un pourcentage croissant de liens temporels. Ainsi, une fraction de plus en plus grande du réseau verrait éliminées ses caractéristiques temporelles. Soient un flot de liens $L = (T, V, E)$, avec $T = [t_\alpha, t_\omega]$, un ensemble vide $E' = \emptyset$ et un nombre décimal $p \in [0, 1]$:

$$\begin{aligned}
&\textit{Pour un nombre } p \times |E| \textit{ de } (t, u, v) \in E, \textit{ on tire } t' \in [t_\alpha, t_\omega] \\
&E' = E' \cup \{(t', u, v)\} \\
&\textit{On remplace } E \textit{ par } E' \textit{ dans } L \textit{ tq } L = (T, V, E')
\end{aligned}$$

- **RT**Δ

Avec ce modèle, tous les liens sont randomisés, mais seulement partiellement : plutôt que de tirer des temps quelconques dans tout l'intervalle de temps des données, on les choisit dans un intervalle réduit. En faisant varier la taille de cet intervalle, on observe la décons-

truction progressive des propriétés ciblées dans le réseau. Ainsi, on observe la sensibilité à la randomisation, c'est-à-dire, à quel point l'intervalle des valeurs de temps doit être large pour que des effets sur la propagation deviennent observables. Soient un flot de liens $L = (T, V, E)$, avec $T = [t_\alpha, t_\omega]$, un ensemble vide $E' = \emptyset$ et un entier $\Delta < t_\omega - t_\alpha$:

$$\forall (t, u, v) \in E,\ \textit{tirage de } t' \in [t - \frac{\Delta}{2}, t + \frac{\Delta}{2}]$$
$$E' = E' \cup \{(t', u, v)\}$$
$$\textit{On remplace } E \textit{ par } E' \textit{ dans } L \textit{ tq } L = (T, V, E')$$

6.5.2 Impact de $RT\Delta$ sur la taille et l'ASCI des cascades

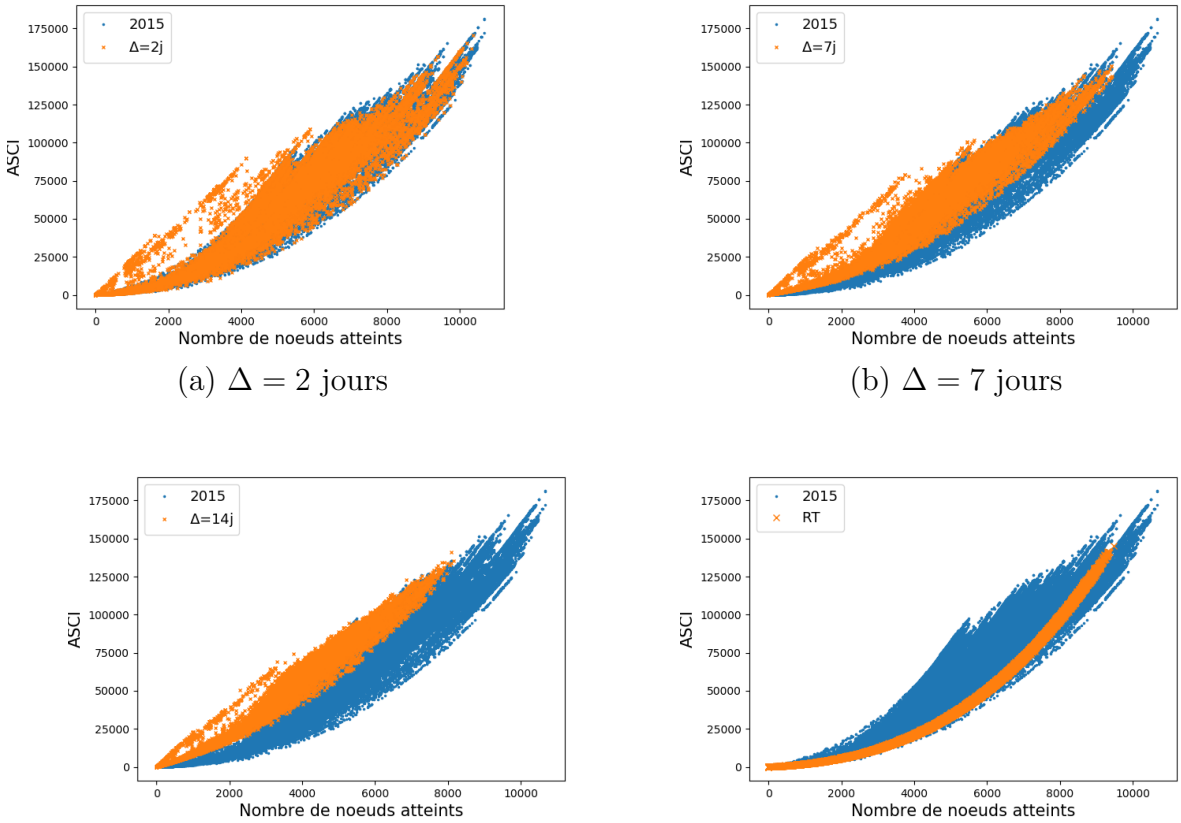
On effectue à présent des mesures exploratoires sur le flot de liens de 2015.

On choisit quatre valeurs de Δ, exprimées en jours : $\Delta = \{0, 2, 7, 14\}$. Une valeur de 0 équivaut à ne pas randomiser les données. On étudie également le réseau randomisé par le modèle RT, qui équivaut à un Δ d'un an. Pour chaque valeur de Δ, on simule des cascades d'un mois au départ d'un échantillon de nœuds, sélectionné aléatoirement. On mesure ensuite la taille des cascades et leurs ASCI. On compare en figure 6.7 la relation entre ces deux variables.

Pour l'échantillon de cascades sélectionné, on remarque que le nombre de nœuds atteints est plus faible lorsque l'on augmente la valeur de Δ. Les caractéristiques temporelles des interactions favoriseraient de plus grandes propagations à cette échelle de temps. On peut par exemple supposer que les corrélations temporelles permettent que les liens temporels s'enchaînent dans le réseau selon un ordre et un temps inter-contact permettant aux propagations d'atteindre plus de nœuds.

Cependant, lorsque l'on regarde les cascades obtenues avec le modèle RT (figure 6.7d), on constate que la destruction totale des caractéristiques temporelles permet d'obtenir des tailles plus grandes que lorsque $\Delta = 14$ jours. Par contre, ces cascades obtiennent des ASCI très similaires entre elles, et de valeurs plutôt faibles comparativement aux résultats sans randomisation ou randomisés selon le modèle $RT\Delta$. Il faudrait tester des valeurs intermédiaires de Δ pour arriver à déterminer la bascule dans le changement de dynamique de diffusion. Il serait également intéressant d'étudier les profils de propagation des cascades obtenues avec chaque valeurs de Δ.

Par ailleurs, nous avons déterminé au chapitre 5 que l'allure en faisceau de paraboles du nuage de points de l'ASCI en fonction du nombre de nœuds atteints dépend des valeurs de pentes des cascades. On remarque sur la figure 6.7 que pour un nombre donné de nœuds atteints, la diversité des valeurs d'ASCI diminue lorsque Δ augmente. On peut ainsi faire



(a) $\Delta = 2$ jours (b) $\Delta = 7$ jours

(c) $\Delta = 14$ jours (d) RT

FIGURE 6.7 – ASCI en fonction des tailles de cascades, pour un échantillon des propagations de 2015, sur des flots de liens de 2015 randomisés selon le modèle $RT\Delta$ et RT.

l'hypothèse que la méthode RT diminue la diversité des valeurs de pente des diffusions.

En outre, dès de faibles valeurs de Δ, on voit changer l'allure d'une partie du nuage de point de la figure 6.7 : pour les valeurs les plus élevées d'ASCI en fonction du nombre de nœuds atteints, on remarque que ces deux grandeurs ne sont plus systématiquement liées par une relation parabolique. Au contraire, on voit que la relation entre les fortes ASCI et le nombre de nœuds atteints semble d'allure linéaire. Le modèle à deux phases, défini au chapitre 5 pour décrire l'accessibilité des nœuds au cours du temps, ne serait donc plus satisfaisant pour modéliser ces cascades. Pour vérifier cette hypothèse, on compare le modèle à deux phases avec les propagations randomisées selon deux différentes valeurs de Δ, en figure 6.8. Comme supposé, on remarque que plus Δ augmente, moins le modèle semble adapté pour retrouver des valeurs d'ASCI en fonction du nombre de nœuds atteints représentatives des propagations mesurées. Ainsi, entre autres caractéristiques du réseau temporel, le modèle de randomisation $RT\Delta$ semble aboutir à la destruction de la corrélation

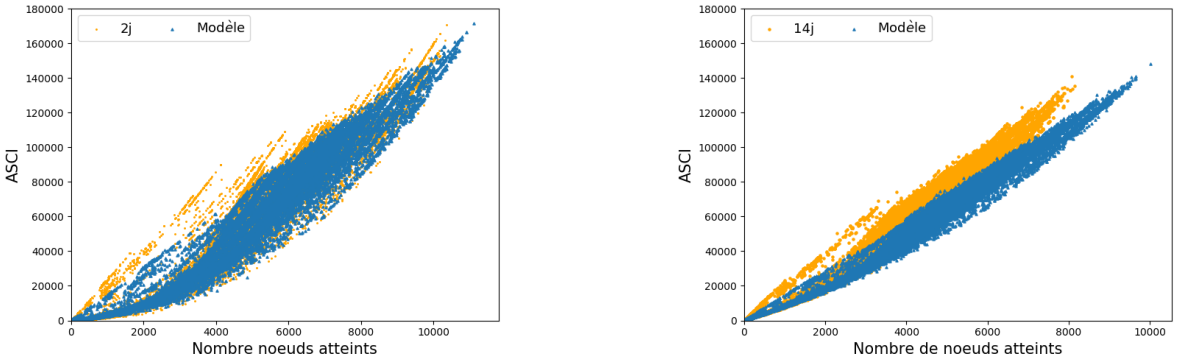


FIGURE 6.8 – Comparaison de l'ASCI en fonction tailles de cascades, obtenues avec le modèle à deux phases (noté modèle) ou sur des propagations au départ d'un cinquième des nœuds de 2015. A gauche, $\Delta = 2j$; à droite, $\Delta = 14j$.

entre les pentes des diffusions et le temps d'attente.

Dans cette partie, nous avons donné des pistes pour étudier des propriétés structurelles et temporelles impactant la dynamique de propagation sur le réseau. En plus d'utiliser de plus grandes valeurs de Δ pour randomiser le réseau avec le modèle $RT\Delta$, nous souhaitons étudier la randomisation progressive du réseau avec le modèle $RT\%$.

Par ailleurs, d'autres modèles de la littérature seraient intéressants pour mener une étude du même type. [Gauvin et al., 2013] proposent un modèle rendant aléatoire les temps inter-contact séparant deux activations d'un même lien, afin d'étudier l'impact de la distribution de ces temps sur la propagation. Comme nous pensons que cette propriété joue un rôle important dans le processus de diffusion, nous pourrions mener une étude analogue en randomisant progressivement les temps inter-contact des données initiales. De plus, dans [Tabourier et al., 2012], les auteurs mettent en valeur le rôle de la causalité entre les liens dans le processus de diffusion : considérant l'ensemble des nœuds atteints par un lien ayant pour origine un nœud donné, le modèle proposé par les auteurs consiste à échanger aléatoirement ces nœuds atteints. Nous pourrions donc randomiser progressivement les données avec ce modèle, afin d'en observer l'impact sur les propagations simulées sur la BDNI.

Chapitre 7

Annexes

Sommaire

7.1 Volumes de suppression

Dans le chapitre 4, nous décrivons la conversion d'un nombre de nœuds (ou de liens statiques) en son équivalent en nombre de liens temporels, nommé volume de suppression (cf définition 4.2). Nous présentons ici les volumes de suppression des stratégies d'identification des nœuds et liens statiques de l'année 2015. Pour rappel (cf tableau 2.1), l'année 2015 comporte :

— 176 771 nœuds actifs ;
— 167 193 nœuds de degré sortant non nul ;
— 101 481 nœuds de degré entrant non nul ;
— 1 037 465 liens statiques ;
— 2 652 131 liens temporels.

Le tableau suivant présente les volumes de suppression (exprimés en nombre de liens temporels) supprimés par les stratégies d'identification des nœuds sur l'année 2015. On note N_{occ} le nombre d'occurrences dans les cascades, k le degré, $\mathcal{A}$ l'activité, BC la centralité d'intermédiarité des nœuds.

Nombre de nœuds	1	10	25	50	75	90	100
k	38852	213344	434717	685413	865089	950393	997892
k_{in}	38852	208566	433887	686274	849437	934112	987157
k_{out}	16066	192472	379140	556956	741164	812970	872793
$\mathcal{A}_{in}$	38852	234241	451195	716867	891965	978292	1030521
$\mathcal{A}_{out}$	27774	175065	411740	629920	753836	860436	908241
BC	27774	189837	375420	612366	788988	854554	907047
$N_{occ}(u)$	16066	186568	351094	612686	777319	872991	922409

La tableau suivant présente les volumes de suppression de la stratégie d'identification des liens statiques selon leur poids w (définition 1.7) mesuré en 2015 :

Nombre de liens statiques	Nombre de liens temporels
140	16 011
4 360	192 478
5 279	218 965
12 741	392 653
45 595	857 552
92 126	1 221 637
25 650	612 277
216 651	1 692 603
421 135	2 015 851

Le tableau suivant présente les volumes de suppression de la stratégie d'identification des liens statiques selon leur nombre d'occurrences dans les cascades (définition 4.1) de

2015 :

Nombre de liens statiques	Nombre de liens temporels
458	16 021
6 409	85 922
48 288	192 471
106 928	392 656
216 651	782 945
410 152	1 592 612
176 000	612 624
165 000	557 606

7.2 Représentation en nœud papillon

On observe que la répartition des nœuds dans les différentes structures du nœud papillon est stable sur les graphes annuels. On remarque une légère diminution de la taille de la composante faiblement connexe géante dans les années 2007 et 2008, qui passe d'environ 98% des nœuds à 90. Ceci peut potentiellement être lié à l'épizootie de fièvre catarrhale ovine, ayant touchée la France ces années là [1]. Bien qu'affectant majoritairement les ovins, cette maladie peut également toucher les bovins. Les restrictions sanitaires découlant de cette crise ont pu impacter les échanges.

1. https://www.anses.fr/fr/content/la-fi%C3%A8vre-catarrhale-ovine-fco-ou-bluetongue

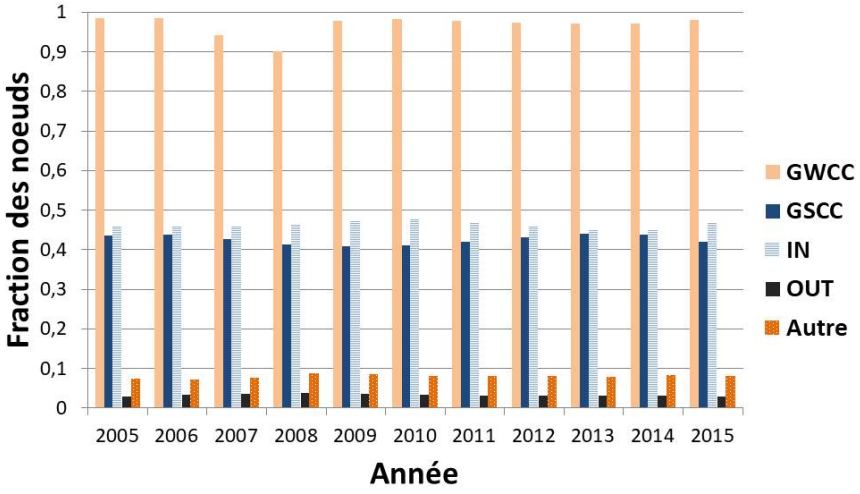
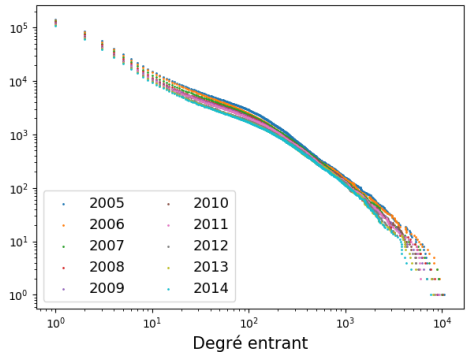


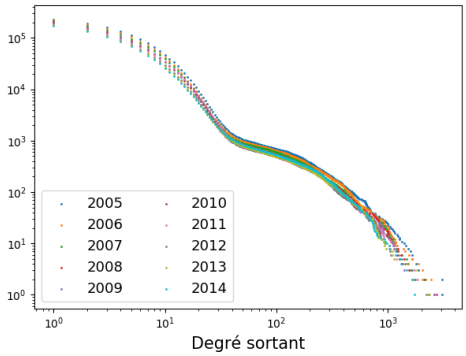
FIGURE 7.1 – Fractions des nœuds de la séquence de graphes annuels, appartenant aux différentes structures du nœud papillon.

7.3 Distributions de degrés

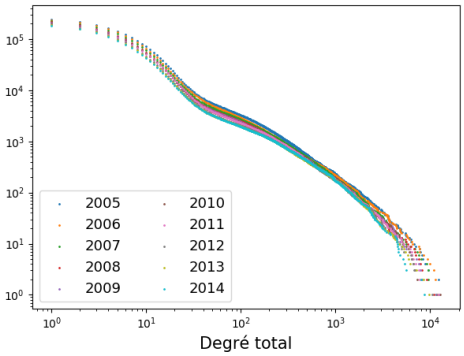
Figure 7.2, on observe que les distributions des degrés entrant, sortant et total ont une allure similaire d'une année sur l'autre. Les valeurs maximales de chaque type de degré sont du même ordre de grandeur d'année en année. On remarque toutefois une tendance à la baisse de la valeur maximale du degré sortant au cours des années (diminue de plus de 40%), alors que cette diminution est plus faible (d'environ 20%) pour la valeur maximale du degré entrant. Cette évolution est probablement liée à la diminution du nombre de nœuds et de liens statiques actifs dans les graphes annuels.



(a) Degré entrant



(b) Degré sortant

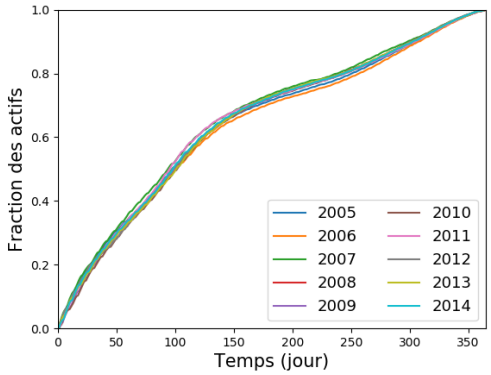


(c) Degré total

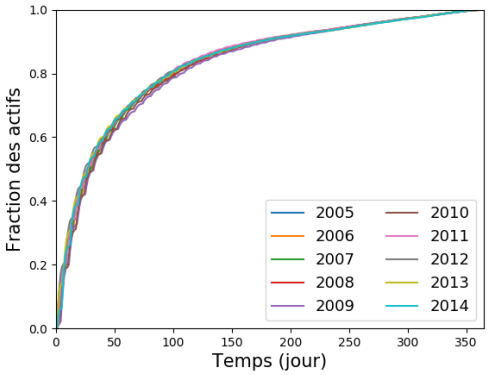
FIGURE 7.2 – Distributions cumulatives inverses des degrés de la séquence de graphes annuels.

7.4 Dynamique d'apparition des nœuds

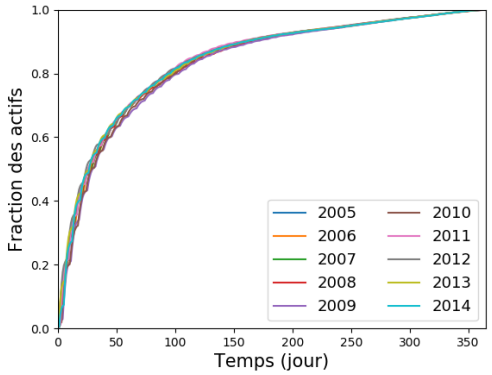
Les conclusions tirées de l'analyse de l'année 2015 (voir partie 2.3.3) se confirment sur les autres années : on observe que les nœuds ayant une activité entrante s'activent tout au long des différentes années, alors que les nœuds ayant une activité sortante s'activent très tôt dans l'année (figure 7.3).



(a) Degré entrant



(b) Degré sortant



(c) Degré total

FIGURE 7.3 – Dynamique d'activation des nœuds achetant des animaux (a), des nœuds vendeurs (b), et des nœuds achetant ou vendant (c). Les résultats sont normalisés respectivement par le nombre de nœuds acheteurs actifs, le nombre de nœuds vendeurs actifs, et le nombre de nœuds actifs total.

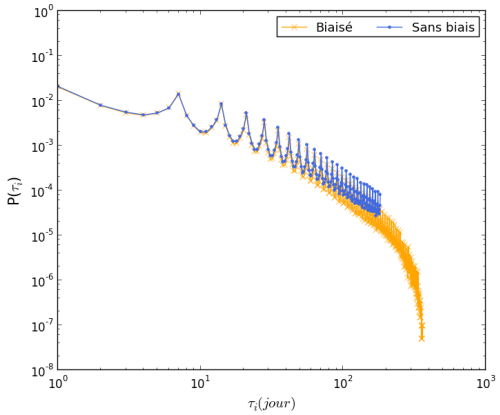
7.5 Distributions des temps inter-contacts

Comme pour la distribution des temps inter-contacts de l'année 2015 (voir partie 2.3.7), on observe sur les flots de liens des années 2005 à 2014 une périodicité des valeurs de temps inter-contact qui correspond aux semaines, et qui s'accompagne d'une décroissance de la probabilité (figures 7.4 et 7.5).

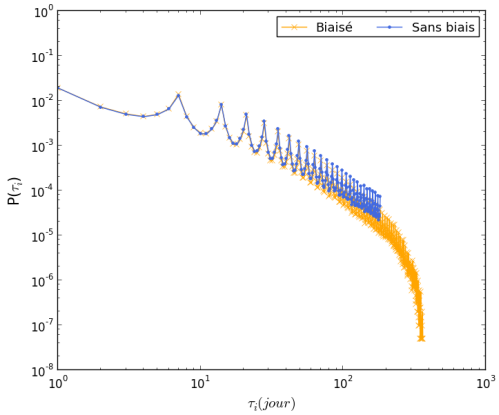
7.6 Comparaison des tailles des cascades sur graphe ou flot de liens

Dans le chapitre 3, nous décrivons un protocole de comparaison des cascades sur graphe et flot de liens. Pour s'assurer que les cascades de durée d'un an n'aient pas toute le même point de départ (le 1er janvier), et qu'elles aient une année entière pour atteindre des nœuds, il faut considérer un flot de liens représentant deux ans de données. Pour que les résultats soient comparables au cas statique, nous prenons donc une séquence de deux graphes annuels pour effectuer les mesures. La figure 7.6 présente les résultats pour des couples d'années, de 2006 à 2015. Nous confirmons les conclusions du chapitre 3 : négliger l'ordre chronologique des interactions fait perdre la diversité des tailles de cascades, et entraîne la surestimation de leur taille.

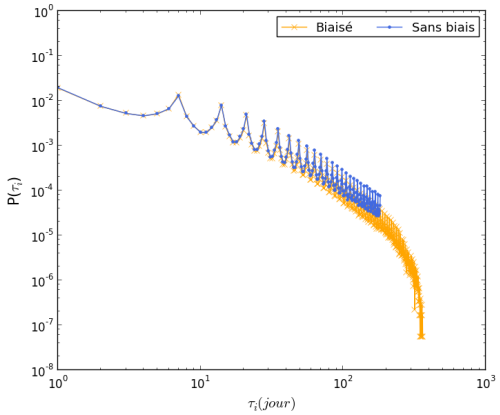
De même, nous simulons des propagations de durée d'un mois et d'un trimestre sur les autres années de la BDNI, pour un échantillon de nœuds de départ. Nous retrouvons les résultats décris sur l'année 2015.



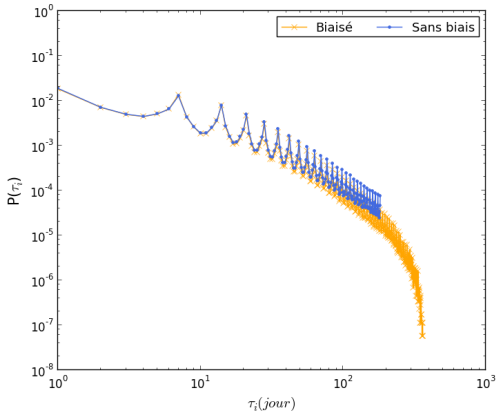
(a) 2005



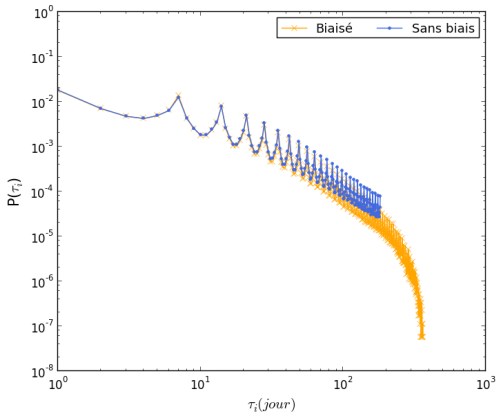
(b) 2006



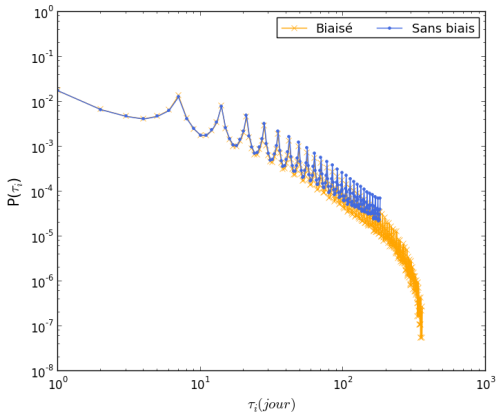
(c) 2007



(d) 2008

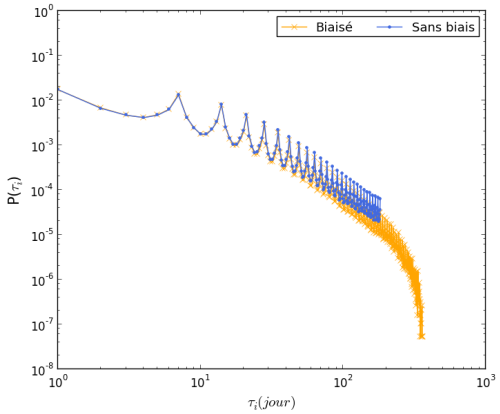


(e) 2009

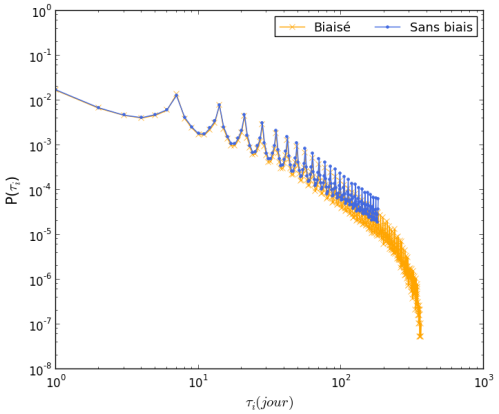


(f) 2010

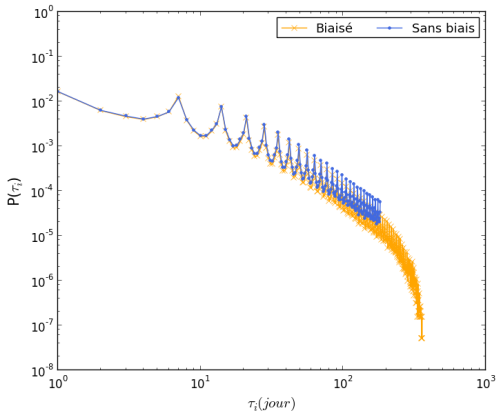
FIGURE 7.4 – Distribution des temps inter-contact entre deux interactions, pour les flots de liens des années 2005 à 2010, avec ou sans correction du biais. Échelles logarithmiques.



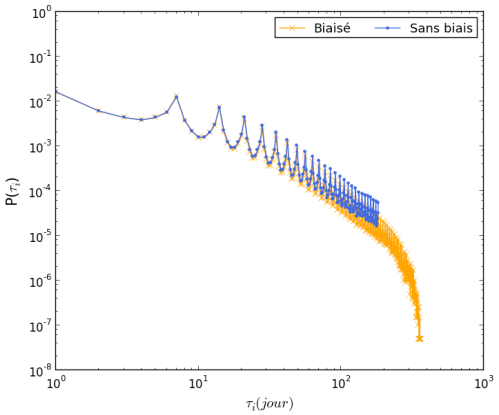
(a) 2011



(b) 2012

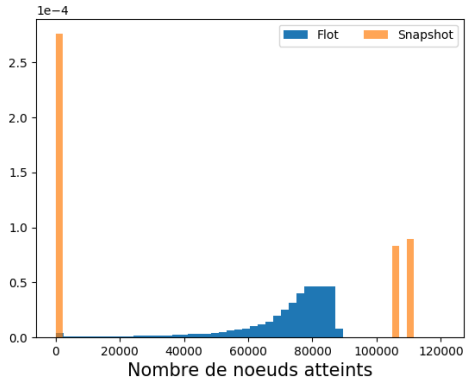


(c) 2013

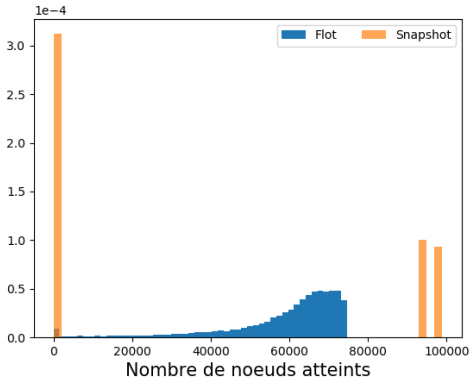


(d) 2014

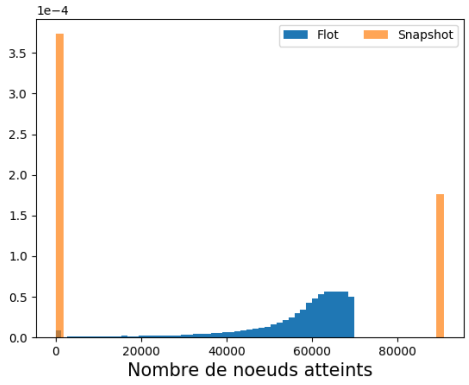
FIGURE 7.5 – Distribution des temps inter-contact entre deux interactions, pour les flots de liens des années 2011 à 2014, avec ou sans correction du biais. Échelles logarithmiques.



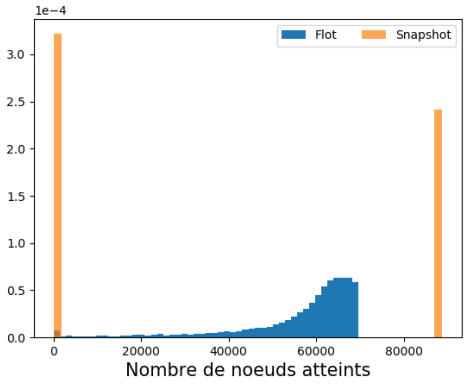
(a) 2006-2007



(b) 2008-2009

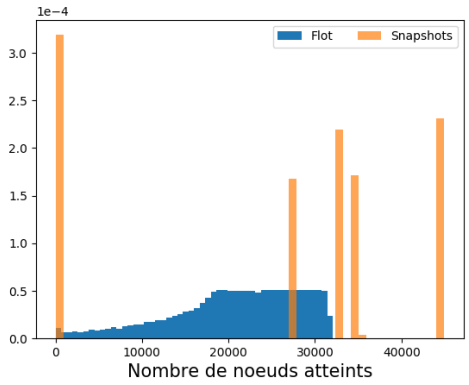


(c) 2010-2011

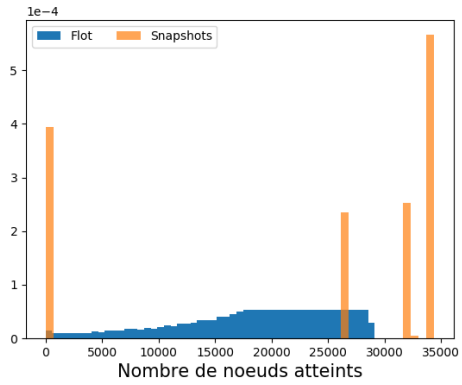


(d) 2012-2013

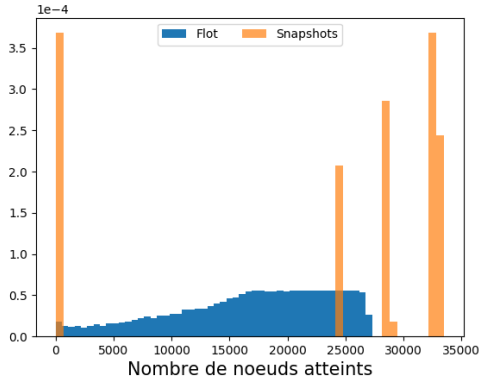
FIGURE 7.6 – Distributions des tailles de cascades de durée d'un an, sur flots de liens et graphes annuels. Valeurs discrétisées, en 50 intervalles.



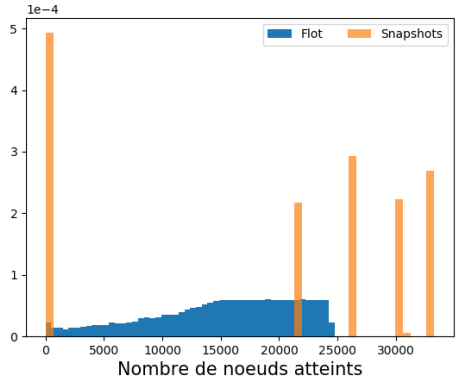
(a) 2006



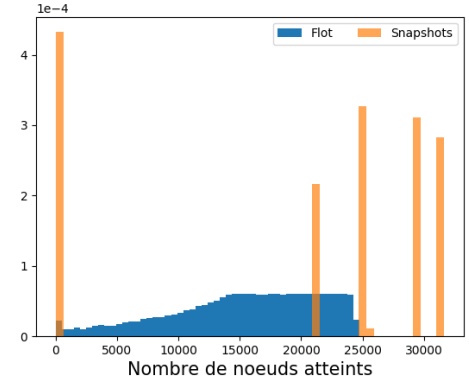
(b) 2007



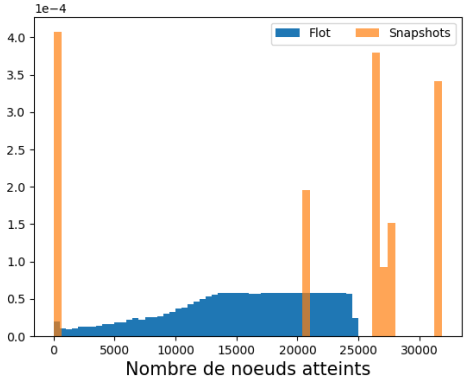
(c) 2008



(d) 2009

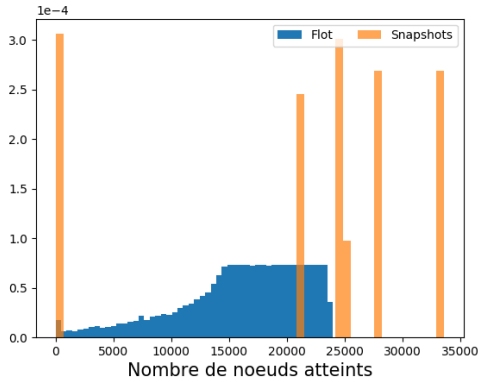


(e) 2010

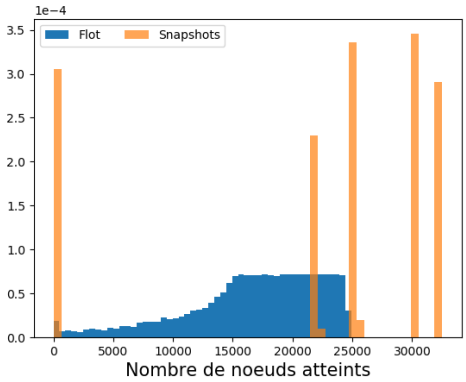


(f) 2011

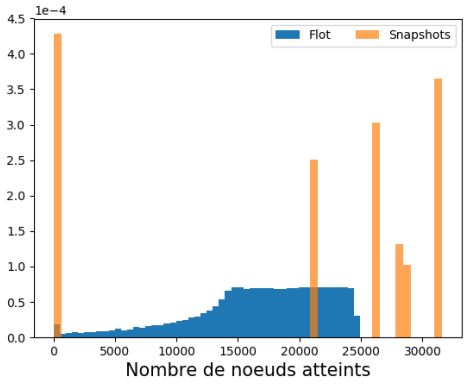
FIGURE 7.7 – Distributions des tailles de cascades de durée d'un trimestre, sur flots de liens et graphes trimestriels. Valeurs discrétisées, en 50 intervalles.



(a) 2012

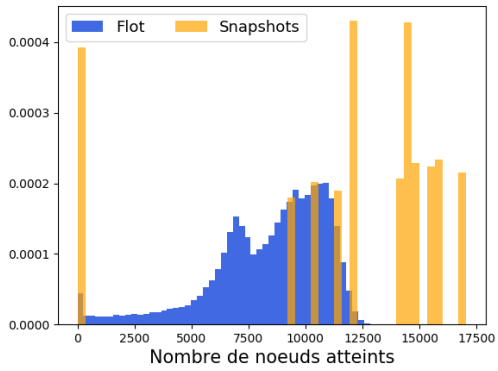


(b) 2013

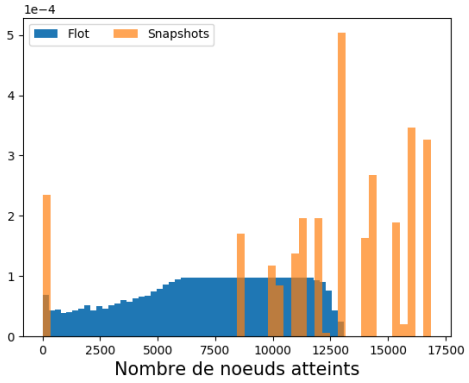


(c) 2014

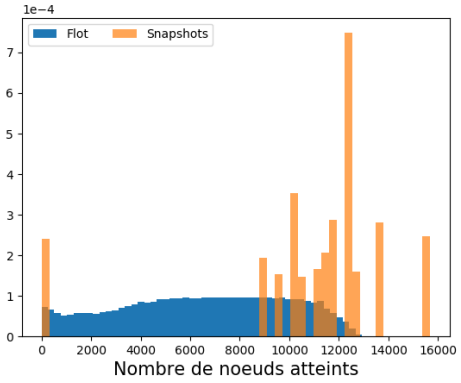
FIGURE 7.8 – Distributions des tailles de cascades de durée d'un trimestre, sur flots de liens et graphes trimestriels. Valeurs discrétisées, en 50 intervalles.



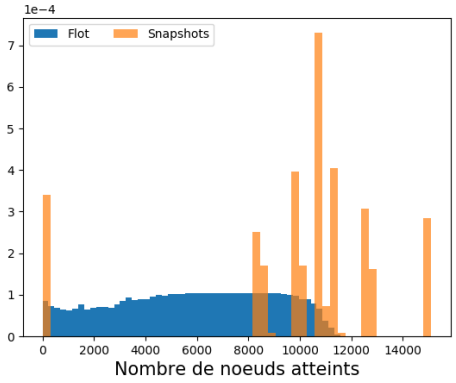
(a) 2005



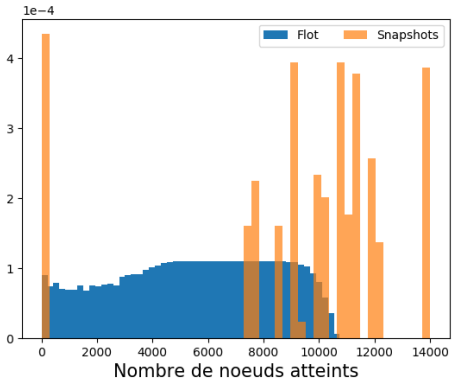
(b) 2006



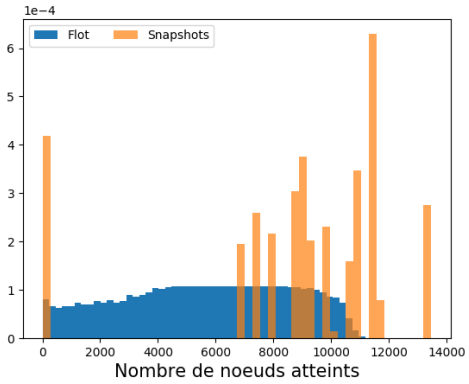
(c) 2007



(d) 2008

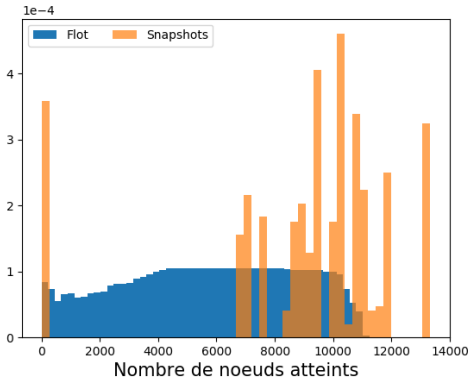


(e) 2009

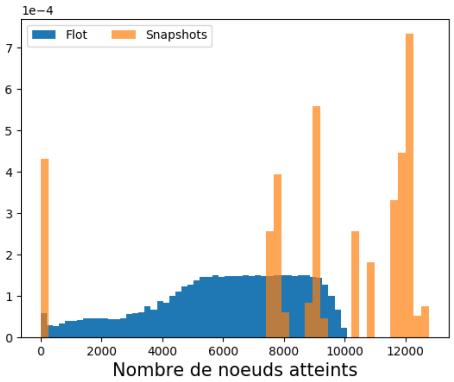


(f) 2010

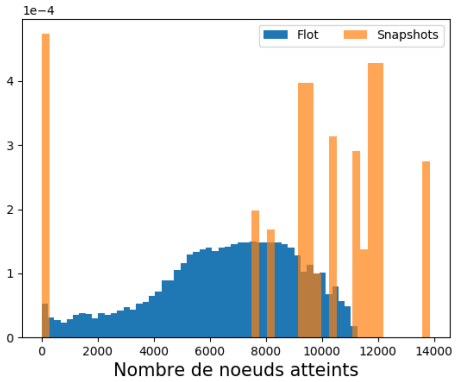
FIGURE 7.9 – Distributions des tailles de cascades de durée d'un mois, sur flots de liens et graphes mensuels. Valeurs discrétisées, en 50 intervalles.



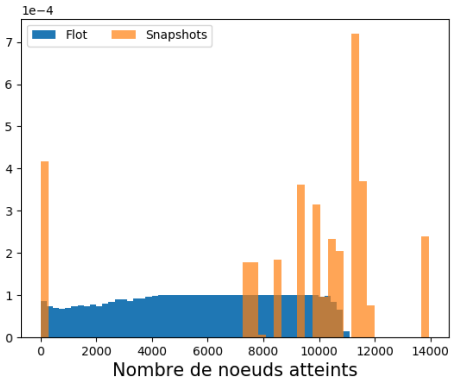
(a) 2011



(b) 2012



(c) 2013



(d) 2014

FIGURE 7.10 – Distributions des tailles de cascades de durée d'un mois, sur flots de liens et graphes mensuels. Valeurs discrétisées, en 50 intervalles.

Bibliographie

[Anderson and May, 1992] Anderson, R. M. and May, R. M. (1992). *Infectious diseases of humans : dynamics and control*. Oxford university press.

[Bajardi et al., 2011] Bajardi, P., Barrat, A., Natale, F., Savini, L., and Colizza, V. (2011). Dynamical patterns of cattle trade movements. *PloS one*, 6(5) :e19869.

[Barrat et al., 2008] Barrat, A., Barthlemy, M., and Vespignani, A. (2008). *Dynamical Processes on Complex Networks*. Cambridge University Press, New York, NY, USA, 1st edition.

[Barthélemy et al., 2004] Barthélemy, M., Barrat, A., Pastor-Satorras, R., and Vespignani, A. (2004). Velocity and hierarchical spread of epidemic outbreaks in scale-free networks. *Physical Review Letters*, 92(17) :178701.

[Bavelas, 1950] Bavelas, A. (1950). Communication patterns in task-oriented groups. *The Journal of the Acoustical Society of America*, 22(6) :725–730.

[Beaunée et al., 2015] Beaunée, G., Vergu, E., and Ezanno, P. (2015). Modelling of paratuberculosis spread between dairy cattle farms at a regional scale. *Veterinary research*, 46(1) :111.

[Belik et al., 2015] Belik, V., Fengler, A., Fiebig, F., Lentz, H. H., and Hövel, P. (2015). Controlling contagious processes on temporal networks via adaptive rewiring. *arXiv preprint arXiv :1509.04054*.

[Bollobás, 1984] Bollobás, B. (1984). *Graph Theory and Combinatorics : Proceedings of the Cambridge Combinatorial Conference in Honour of Paul Erdös,[Trinity College, Cambridge, 21-25 March 1983]*. Academic Press.

[Brandes and Pich, 2007] Brandes, U. and Pich, C. (2007). Centrality estimation in large networks. *International Journal of Bifurcation and Chaos*, 17(07) :2303–2318.

[Brin and Page, 1998] Brin, S. and Page, L. (1998). The anatomy of a large scale hypertextual web search engine, computer networks and isdn systems 30 (1-7).

[Broder et al., 2000] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000). Graph structure in the web. *Computer networks*, 33(1-6) :309–320.

[Brooks-Pollock et al., 2014] Brooks-Pollock, E., Roberts, G. O., and Keeling, M. J. (2014). A dynamic model of bovine tuberculosis spread and control in great britain. *Nature*, 511(7508) :228.

[Büttner et al., 2013] Büttner, K., Krieter, J., Traulsen, A., and Traulsen, I. (2013). Efficient interruption of infection chains by targeted removal of central holdings in an animal trade network. *PLoS One*, 8(9) :e74292.

[Büttner et al., 2016] Büttner, K., Krieter, J., Traulsen, A., and Traulsen, I. (2016). Epidemic spreading in an animal trade network–comparison of distance-based and network-based control measures. *Transboundary and emerging diseases*, 63(1).

[Cameron, 2012] Cameron, A. (2012). The consequences of risk-based surveillance : Developing output-based standards for surveillance to demonstrate freedom from disease. *Preventive veterinary medicine*, 105(4) :280–286.

[Casteigts et al., 2012] Casteigts, A., Flocchini, P., Quattrociocchi, W., and Santoro, N. (2012). Time-varying graphs and dynamic networks. *International Journal of Parallel, Emergent and Distributed Systems*, 27(5) :387–408.

[Chen et al., 2008] Chen, Y., Paul, G., Havlin, S., Liljeros, F., and Stanley, H. E. (2008). Finding a better immunization strategy. *Physical review letters*, 101(5) :058701.

[Christley et al., 2005] Christley, R., Robinson, S., Lysons, R., and French, N. (2005). Network analysis of cattle movement in great britain. *Proc. Soc. Vet. Epidemiol. Prev. Med*, pages 234–243.

[Cohen et al., 2003] Cohen, R., Havlin, S., and Ben-Avraham, D. (2003). Efficient immunization strategies for computer networks and populations. *Physical review letters*, 91(24) :247901.

[Courcoul and Ezanno, 2010] Courcoul, A. and Ezanno, P. (2010). Modelling the spread of bovine viral diarrhoea virus (bvdv) in a managed metapopulation of cattle herds. *Veterinary microbiology*, 142(1-2) :119–128.

[Crépey and Barthélemy, 2007] Crépey, P. and Barthélemy, M. (2007). Detecting robust patterns in the spread of epidemics : a case study of influenza in the united states and france. *American journal of epidemiology*, 166(11) :1244–1251.

[Direction générale de la mondialisation, 2011] Direction générale de la mondialisation, du développement, e. d. p. (2011). Position française sur le concept « one health/une seule santé ».

[Dubé et al., 2008] Dubé, C., Ribble, C., Kelton, D., and McNab, B. (2008). Comparing network analysis measures to determine potential epidemic size of highly contagious exotic diseases in fragmented monthly networks of dairy cattle movements in ontario, canada. *Transboundary and emerging diseases*, 55(9-10) :382–392.

[Dubé et al., 2009] Dubé, C., Ribble, C., Kelton, D., and McNab, B. (2009). A review of network analysis terminology and its application to foot-and-mouth disease modelling and policy development. *Transboundary and emerging diseases*, 56(3) :73–85.

[Dutta et al., 2014] Dutta, B. L., Ezanno, P., and Vergu, E. (2014). Characteristics of the spatio-temporal network of cattle movements in france over a 5-year period. *Preventive veterinary medicine*, 117(1) :79–94.

[Ensoy et al., 2014] Ensoy, C., Faes, C., Welby, S., Van der Stede, Y., and Aerts, M. (2014). Exploring cattle movements in belgium. *Preventive veterinary medicine*, 116(1-2) :89–101.

[EU, 2000] EU (2000). Regulation (ec) no 1760/2000 of the european parliament and of the council of 17 july 2000 establishing a system for the identification and registration of bovine animals. OJ L 204, 11.8.2000, p. 1–10.

[Eubank et al., 2004] Eubank, S., Guclu, H., Kumar, V. A., Marathe, M. V., Srinivasan, A., Toroczkai, Z., and Wang, N. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988) :180.

[Ezanno et al., 2007] Ezanno, P., Fourichon, C., Viet, A.-F., and Seegers, H. (2007). Sensitivity analysis to identify key-parameters in modelling the spread of bovine viral diarrhoea virus in a dairy herd. *Preventive veterinary medicine*, 80(1) :49–64.

[Fournet and Barrat, 2014] Fournet, J. and Barrat, A. (2014). Contact patterns among high school students. *PloS one*, 9(9) :e107878.

[FranceAgriMer, 2012] FranceAgriMer (2012). Le commerce international de la viande bovine, vers une stabilisation des échanges ? Issue Numéro 16 élevages/viande.

[Freeman, 1977] Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41.

[Gauvin et al., 2018] Gauvin, L., Génois, M., Karsai, M., Kivelä, M., Takaguchi, T., Valdano, E., and Vestergaard, C. L. (2018). Randomized reference models for temporal networks. *arXiv preprint arXiv :1806.04032*.

[Gauvin et al., 2013] Gauvin, L., Panisson, A., Cattuto, C., and Barrat, A. (2013). Activity clocks : spreading dynamics on temporal networks of human contact. *Sci. Rep.*, 3 :3099. 17 pages, 6 figures.

[Génois et al., 2015] Génois, M., Vestergaard, C. L., Cattuto, C., and Barrat, A. (2015). Compensating for population sampling in simulations of epidemic spread on temporal contact networks. *Nature communications*, 6 :8860.

[Green et al., 2006] Green, D., Kiss, I., and Kao, R. (2006). Modelling the initial spread of foot-and-mouth disease through animal movements. *Proceedings of the Royal Society of London B : Biological Sciences*, 273(1602) :2729–2735.

[Harvey et al., 2007] Harvey, N., Reeves, A., Schoenbaum, M. A., Zagmutt-Vergara, F. J., Dubé, C., Hill, A. E., Corso, B. A., McNab, W. B., Cartwright, C. I., and Salman, M. D. (2007). The north american animal disease spread model : A simulation model to assist decision making in evaluating animal disease incursions. *Preventive veterinary medicine*, 82(3-4) :176–197.

[Holme and Saramäki, 2012] Holme, P. and Saramäki, J. (2012). Temporal networks. *Physics reports*, 519(3) :97–125.

[Isella et al., 2011] Isella, L., Stehlé, J., Barrat, A., Cattuto, C., Pinton, J.-F., and Van den Broeck, W. (2011). What's in a crowd ? analysis of face-to-face behavioral networks. *Journal of theoretical biology*, 271(1) :166–180.

[Kao et al., 2006] Kao, R., Danon, L., Green, D., and Kiss, I. (2006). Demographic structure and pathogen dynamics on the network of livestock movements in great britain. *Proceedings of the Royal Society of London B : Biological Sciences*, 273(1597) :1999–2007.

[Karsai et al., 2011] Karsai, M., Kivelä, M., Pan, R. K., Kaski, K., Kertész, J., Barabási, A.-L., and Saramäki, J. (2011). Small but slow world : How network topology and burstiness slow down spreading. *Physical Review E*, 83(2) :025102.

[Katz, 1953] Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1) :39–43.

[Keeling and Eames, 2005] Keeling, M. J. and Eames, K. T. (2005). Networks and epidemic models. *Journal of the Royal Society Interface*, 2(4) :295–307.

[Kiss et al., 2006] Kiss, I. Z., Green, D. M., and Kao, R. R. (2006). The network of sheep movements within great britain : network properties and their implications for infectious disease spread. *Journal of the Royal Society Interface*, 3(10) :669–677.

[Kitsak et al., 2010] Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., and Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Nature physics*, 6(11) :888.

[Kostakos, 2009] Kostakos, V. (2009). Temporal graphs. *Physica A : Statistical Mechanics and its Applications*, 388(6) :1007–1023.

[Kovanen et al., 2011] Kovanen, L., Karsai, M., Kaski, K., Kertész, J., and Saramäki, J. (2011). Temporal motifs in time-dependent networks. *Journal of Statistical Mechanics : Theory and Experiment*, 2011(11) :P11005.

[Lambiotte et al., 2013] Lambiotte, R., Tabourier, L., and Delvenne, J.-C. (2013). Burstiness and spreading on temporal networks. *The European Physical Journal B*, 86(7) :320.

[Latapy et al., 2017] Latapy, M., Viard, T., and Magnien, C. (2017). Stream graphs and link streams for the modeling of interactions over time. *arXiv preprint arXiv :1710.04073*.

[Le Menach et al., 2006] Le Menach, A., Vergu, E., Grais, R. F., Smith, D. L., and Flahault, A. (2006). Key strategies for reducing spread of avian influenza among commercial poultry holdings : lessons for transmission to humans. *Proceedings of the Royal Society of London B : Biological Sciences*, 273(1600) :2467–2475.

[Lee et al., 2016] Lee, M.-J., Choi, S., and Chung, C.-W. (2016). Efficient algorithms for updating betweenness centrality in fully dynamic graphs. *Information Sciences*, 326 :278–296.

[Lentz et al., 2016] Lentz, H. H., Koher, A., Hövel, P., Gethmann, J., Sauter-Louis, C., Selhorst, T., and Conraths, F. J. (2016). Disease spread through animal movements : a static and temporal network analysis of pig trade in germany. *PloS one*, 11(5) :e0155196.

[Liljeros et al., 2001] Liljeros, F., Edling, C. R., Amaral, L. A. N., Stanley, H. E., and Åberg, Y. (2001). The web of human sexual contacts. *Nature*, 411(6840) :907.

[Lipton et al., 1979] Lipton, R. J., Rose, D. J., and Tarjan, R. E. (1979). Generalized nested dissection. *SIAM journal on numerical analysis*, 16(2) :346–358.

[Magnien et al., 2011] Magnien, C., Latapy, M., and Guillaume, J.-L. (2011). Impact of random failures and attacks on poisson and power-law random networks. *ACM Computing Surveys (CSUR)*, 43(3) :13.

[Ministère de l'agriculture et de l'alimentation, 2004] Ministère de l'agriculture et de l'alimentation, d. g. d. l. (2004). Enregistrement des exploitations et des détenteurs dans le cadre de l'identification et de la traçabilité des animaux d'élevage.

[Miritello et al., 2011] Miritello, G., Moro, E., and Lara, R. (2011). Dynamical strength of social ties in information spreading. *Physical Review E*, 83(4) :045102.

[Moreno et al., 2004] Moreno, Y., Nekovee, M., and Pacheco, A. F. (2004). Dynamics of rumor spreading in complex networks. *Physical Review E*, 69(6) :066130.

[Morris and Kretzschmar, 1997] Morris, M. and Kretzschmar, M. (1997). Concurrent partnerships and the spread of hiv. *Aids*, 11(5) :641–648.

[Mweu et al., 2013] Mweu, M. M., Fournié, G., Halasa, T., Toft, N., and Nielsen, S. S. (2013). Temporal characterisation of the network of danish cattle movements and its implication for disease control : 2000–2009. *Preventive veterinary medicine*, 110(3-4) :379–387.

[Newman, 2010] Newman, M. (2010). *Networks : an introduction*. Oxford university press.

[Nöremark et al., 2011] Nöremark, M., Håkansson, N., Lewerin, S. S., Lindberg, A., and Jonsson, A. (2011). Network analysis of cattle and pig movements in sweden : measures relevant for disease control and risk based surveillance. *Preventive veterinary medicine*, 99(2-4) :78–90.

[Onnela et al., 2007] Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A.-L. (2007). Structure and tie strengths in mobile communication networks. *Proceedings of the national academy of sciences*, 104(18) :7332–7336.

[Pastor-Satorras and Vespignani, 2002] Pastor-Satorras, R. and Vespignani, A. (2002). Immunization of complex networks. *Physical Review E*, 65(3) :036104.

[Peruani and Tabourier, 2011] Peruani, F. and Tabourier, L. (2011). Directedness of information flow in mobile phone communication networks. *PloS one*, 6(12) :e28860.

[Rahman et al., 2018] Rahman, M., Saha, T. K., Hasan, M. A., Xu, K. S., and Reddy, C. K. (2018). Dylink2vec : Effective feature representation for link prediction in dynamic networks. *CoRR*, abs/1804.05755.

[Rautureau, 2012] Rautureau, S. (2012). *Simulations d'épizooties de fièvre aphteuse et aide à la décision : approches épidémiologique et économique*. PhD thesis, Université Paris Sud - Paris XI.

[Rautureau et al., 2011] Rautureau, S., Dufour, B., and Durand, B. (2011). Vulnerability of animal trade networks to the spread of infectious diseases : a methodological approach applied to evaluation and emergency control strategies in cattle, france, 2005. *Transboundary and emerging diseases*, 58(2) :110–120.

[Robinson et al., 2007] Robinson, S., Everett, M., and Christley, R. (2007). Recent network evolution increases the potential for large epidemics in the british cattle population. *Journal of the Royal Society Interface*, 4(15) :669–674.

[Rocha and Blondel, 2012] Rocha, L. E. C. and Blondel, V. D. (2012). Temporal heterogeneities increase the prevalence of epidemics on evolving networks. *arXiv preprint arXiv :1206.6036*.

[Roselli et al., 2000] Roselli, D. S., Lorch, J. R., Anderson, T. E., et al. (2000). A comparison of file system workloads. In *USENIX annual technical conference, general track*, pages 41–54.

[Salathé and Jones, 2010] Salathé, M. and Jones, J. H. (2010). Dynamics and control of diseases in networks with community structure. *PLoS computational biology*, 6(4) :e1000736.

[Santoro et al., 2011] Santoro, N., Quattrociocchi, W., Flocchini, P., Casteigts, A., and Amblard, F. (2011). Time-varying graphs and social network analysis : Temporal indicators and metrics. *arXiv preprint arXiv :1102.0629*.

[Schärrer et al., 2015] Schärrer, S., Widgren, S., Schwermer, H., Lindberg, A., Vidondo, B., Zinsstag, J., and Reist, M. (2015). Evaluation of farm-level parameters derived from animal movements for use in risk-based surveillance programmes of cattle in switzerland. *BMC veterinary research*, 11(1) :149.

[Scholtes et al., 2014] Scholtes, I., Wider, N., Pfitzner, R., Garas, A., Tessone, C. J., and Schweitzer, F. (2014). Causality-driven slow-down and speed-up of diffusion in non-markovian temporal networks. *Nature communications*, 5 :5024.

[Starnini et al., 2013] Starnini, M., Machens, A., Cattuto, C., Barrat, A., and Pastor-Satorras, R. (2013). Immunization strategies for epidemic processes in time-varying contact networks. *Journal of theoretical biology*, 337 :89–100.

[Tabourier et al., 2012] Tabourier, L., Stoica, A., and Peruani, F. (2012). How to detect causality effects on large dynamical communication networks : A case study. pages 1–7.

[Tang et al., 2010] Tang, J., Musolesi, M., Mascolo, C., Latora, V., and Nicosia, V. (2010). Analysing information flows and key mediators through temporal centrality metrics. In *Proceedings of the 3rd Workshop on Social Network Systems*, page 3. ACM.

[Vazquez et al., 2007] Vazquez, A., Racz, B., Lukacs, A., and Barabasi, A.-L. (2007). Impact of non-poissonian activity patterns on spreading processes. *Physical review letters*, 98(15) :158702.

[Vernon and Keeling, 2009] Vernon, M. C. and Keeling, M. J. (2009). Representing the uk's cattle herd as static and dynamic networks. *Proceedings of the Royal Society of London B : Biological Sciences*, 276(1656) :469–476.

[Vestergaard et al., 2016] Vestergaard, C. L., Valdano, E., Génois, M., Poletto, C., Colizza, V., and Barrat, A. (2016). Impact of spatially constrained sampling of temporal contact networks on the evaluation of the epidemic risk. *European Journal of Applied Mathematics*, 27(6) :941–957.

[Wietrzyk and Radenkovic, 2007] Wietrzyk, B. and Radenkovic, M. (2007). CRAWDAD dataset nottingham/cattle (v. 2007-12-20). Downloaded from `https://crawdad.org/nottingham/cattle/20071220`.